Fall 12-2011

Identification of Soybean Seed Oil QTLs with Little or No Impact on Seed Protein

Yu-Kai Sun
University of Nebraska-Lincoln

IDENTIFICATION OF SOYBEAN SEED OIL QTLS WITH LITTLE OR NO IMPACT ON SEED PROTEIN

By

Yu-Kai Sun

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the degree of Master of Science

Major: Agronomy

Under the Supervision of Professor James E. Specht

Lincoln, Nebraska

December, 2011

IDENTIFICATION OF SOYBEAN SEED OIL QTLS

WITH LITTLE OR NO IMPACT ON SEED PROTEIN

Yu-Kai Sun, M.S.

University of Nebraska, 2011

Advisor: James E. Specht

A QTL (Quantitative Trait Locus) is chromosomal location of a gene controlling a specific phenotypic characteristic (trait). This trait might be governed by two or more genes and may be affected by environmental interaction. The USA soybean seed composition, when averaged over years and states, is 18.7% oil and 35.3% protein. Soybean seed provides cooking oil for humans and protein for livestock. Concurrent genetic improvement of seed protein (pro) and oil content has been difficult to achieve due to the negative genetic correlation of the two traits. This negative correlation could be due to a pair of tightly linked protein and oil QTLs, whose individual alleles are linkage-paired to give rise to high pro - low oil or low pro – high oil, phenotypes, OR it could be due to just one pleiotropic QTL, whose two alleles have inverse effects on both oil and protein. This thesis objective is to find oil QTLs with minimal effect on protein. Three F_2 populations were developed by mating of two high oil lines with each other and with Williams 82, a current high-yield cultivar. About 500 individual F_2 plants in each population produced $F_{2.3}$ seed progenies and then $F_{2.4}$ seed progenies that were phenotyped for seed protein and oil content. Selective genotyping was used to genotype

F_2 plant progenitors of only the highest and the lowest seed oil deciles of $F_{2.4}$ seed progenies. A 1536 SNP locus assay chip was used for genotyping. In the three mapping populations, eight seed oil QTLs with LOD scores greater than 3.0 were detected and mapped on seven linkage groups using R/qtl software. Six statistically significant seed oil QTLs on LG-C2 (Chr6), LG-M (Chr7), LG-B1 (Chr11), LG-F (Chr13), LG-E (Chr15), and LG-L (Chr19) were detected using genome-wide permutation tests ($\alpha = 0.05$). Of the seed oil QTLs detected in this study, only the seed oil QTLs on LG-F (Chr13) have no significant impact on seed protein content. For improving the seed oil content in high yielding soybean cultivars, S17276 allele (Chr13) from the parental line RMLPC1-311-128-128 may be useful to soybean breeders to improve soybean seed oil content without effecting on seed protein content.

ACKNOWLEDGEMENTS

I would like to thank Dr. James E. Specht for giving me the opportunity to learn and work in his lab and his patient and attentive supervision of this research. I would like to express my appreciation to the other members of my M.S. committee: Dr. George Graef, and Dr. Ismail Dweikat. They provide advice and assistance to this research.

I would like to thank Mike Livingston and Paul Nabity for solving technical problems and helping me with a lot of field experiments. I would like to recognize Leslie Korte, Aaron Hoagland, Travis Wegner, and all students working in the seed lab for their assistance with field and greenhouse work.

I would like to thank Dr. Perry Cregan and Dr. David Hyten for the SNP marker analysis at the Soybean Genomics and Improvement Lab, Beltsville Agricultural Research Center-West, USDA.

I would like to thank Luis Posadas for helping me with experimental population development and F_1 SSR confirmation. He also gave me lots of advice when I encountered problems.

I would like to thank my friend Joseph Jedlicka for helping me throughout the whole research. I also would like to thank Dr. piyaporn Phansak and Kyle Kocak for providing advice and knowledge to this research.

Finally, I thank my family for giving me all the support and courage I needed to finish my M.S. study. Without their love and faith, I could have not been able to accomplish my research.

TABLE OF CONTENTS

PAGE

LIST OF ABBREVIATIONS

a – additive effect

AFLP – amplified fragment length polymorphism

bp – base pair

cDNA – complementary deoxyribonucleic acid

Chr–chromosome

C.I. – confidence interval

cM– centriMorgans

d– dominance effect

DNA– deoxyribonucleic acid

EM–expectation maximization algorithm

GRIN– germplasm resources information network

LOD– logarithm of odds

LG– linkage group

MAS– marker-assisted selection

MR– marker regression

NGRP– national genetic resources program

NIR– near-infrared reflectance

PCR– polymerase chain reaction

QTL or QTLs– quantitative trait locus or loci

RAPD– random amplified polymorphic DNA

RFLP– restriction fragment length polymorphism

List of Abbreviations, continued

RIL or RILs– recombinant inbred line or lines

SNP– single nucleotide polymorphism

SSR– simple sequence repeat

STS– sequence-tagged site

USLP– universal soy linkage panel

LIST OF TABLES

List of Tables, Continued

Appendix Table:

LIST OF FIGURES

List of Figures, Continued

Appendix Figures:

List of Figures, Continued

INTRODUCTION

Soybean [*Glycine max* (L.) Merr.] is one of the major crops of the world. A soybean seed contains about 40% protein and about 20% oil. It is primarily used for producing cooking oil for humans and human food protein products as well as a protein meal animal feed stock. More recently, soybean seed oil is used as a source of biodiesel of industrial materials.

In 2010, soybean represented 58% of world oilseed production (SoyStat, 2011). In the same year, the U.S. produced 35% of the world soybean crop which was 3,329 million bushels, and the total soybean crop value that exceeded 38.9 billion dollars. Soybean seed oil production also represented 29% of the world vegetable oil consumption in 2010.

Scientists and breeders have attempted to increase soybean seed protein and seed oil using different methods. Unfortunately, most of the experimental results have shown that there is a negative correlation between soybean seed oil content and protein content (Ramteke et al., 2010). Schwender et al. (2003) and Chung et al. (2003) examined the relationship between oil and protein content and noted that increasing seed protein content by 2% typically results in a simultaneous 1% decrease of seed oil content. This strong inverse correlation between oil and protein content makes it difficult to improve both traits simultaneously.

Soybean seed oil content is known to be a heritable quantitative trait (Burton, 1987). The oil content of a mature seed is affected by genotype, environment, and their interaction. Recently, many scientists have focused on studying quantitative trait loci

(QTLs), which are statistically determined locations in linkage groups (i.e., chromosomes) of putative genes controlling a quantitative trait of interest. Much of the past and present QTL research has focused on seed protein content, with little research directly focused on the detection of seed oil QTLs *per se*.

There is no clear convincing evidence for discerning, amongst the protein QTLs detected to date, whether the negative correlation between soybean protein and oil content is due : (i) to a single pleiotropic QTL with two alleles, wherein one allele simultaneously causes high oil and low protein, and the other allele simultaneously causes low oil and high protein, or (ii) to two tightly linked QTLs, with the high oil allele at the oil QTL and the low protein allele of the protein QTL currently locked into a repulsion phase not yet reversed by recombination event. To date, there also has been no confirmed report of modifier gene that ameliorates the pleiotropic inverse effects, or a recombination event that has created a heretofore recognized coupled linkage phase of high protein QTL allele with a high oil QTL allele.

The objective of this research was to identify QTLs whose alleles have significant additive effects on seed oil content, but have little or no inverse pleiotropic impact on repulsed linkage phase impact on seed protein content. By using the selective genotyping method of QTL detection and matings of high x low seed oil parents, the highest and lowest deciles of $F_{2:4}$ seed oil distribution would be genotyped with 1536 SNP markers. A skewing of the parental SNP allele frequencies in opposite directions in the decile fractions would be indicative of linkage of the SNP locus with a seed oil gene locus (i.e., QTL). If seed oil QTLs are found that do not alter seed protein content significantly, then the high oil alleles at those QTLs could be used to improve soybean seed oil content of

high-yield cultivars whose oil can be used for biodiesel production, but whose meal is not substantively lower in protein. If so, then the markers flanking the high oil allele at that QTL can then be used for marker-assisted introgression of the high oil allele into existing high yield cultivars.

LITERATURE REVIEW

Oil Improvement in Soybean

Soybean is one of the main sources of edible and industrial oil. It was the number one crop relative to oil production in the world in 2010. Soybean accounted for 58% of oilseed production and ranked number two in world with regard to a 29% share of the vegetable oil consumption (Soy Stats, 2011). In the United States in 2010, the soybean oil production was 19 billion pounds and the total soybean oil consumption was 16.6 billion pounds. In that same year, soybean oil accounted for 68% of the U.S. edible fats and oil consumption and 315 million gallons of biodiesel (Soy Stats, 2011). These data clearly show the domestic importance of the value of soybean oil. Improvement in soybean seed oil would not only benefit food oil production but also the biodiesel industry.

The heritability of soybean seed oil is relatively high, especially if the difference of two parental lines is extremely large (Brim, 1973; Burton, 1987). Chung et al. (2003) reported that the heritability of seed oil content was 0.84. There have been many research studies that have focused on improving soybean seed oil content by simple selection. Panthee et al. (2005) reported that, according to the heritability observed in their study, simple selection should succeed for accomplishing genetic improvement in seed oil. In that study, the observed heritability indicated that genetic variation in seed oil accounted for much of the phenotypic variation. The population mean of a high oil x low oil population tended to approximate the midparent value, leading Thorne and Fehr (1970) to suggest that seed oil is mainly controlled by additive effects instead of dominant effects.

According to the USDA soybean germplasm collection statistics reported in 2001, seed protein ranged from 347 to 552 g kg^{-1} and the seed oil ranged from 65 to 287 g kg^{-1} on a dry seed basis (NGRP, 2001). However, these wide ranges of genetic variability for each trait cannot be taken to indicate that high values of each trait are compatible with each other. In fact, only 15% of all the germplasm accessions had a seed protein content *and* a seed oil content at or above the respective population means for these two traits (Thompson et al., 2001; Chung et al., 2003). This is a reflection of the highly negative correlation between seed protein and oil content, which has been repeatedly reported in many published papers (Hanson et al., 1961; Thorne and Fehr, 1970; Shannon et al., 1972; Brim and Burton, 1979; Sebern and Lambert, 1984; Wehrmann et al., 1987; Hartwig and Kilen, 1991; Helms and Orf, 1998; Cober and Voldeng, 2000). Hanson et al. (1961) suggested that protein/oil conversion ratio was -1.92 when seed yield was constant (i.e., factored out). Schwender et al. (2003) also reached the same conclusion that an increase of 2% in seed protein content typically results in a decrease of 1% of oil content. Chung et al. (2003) reported that the oil-yield phenotypic linear regression coefficient was positive while the protein-yield phenotypic linear regression was negative. These authors also noted that with each yield increase of 1.0 Mg ha^{-1}, oil content increased 1.55 to 1.62 percentage points, while protein content would decreased by 2.34 to 2.86 percentage points. In other words, 1.51 to 1.77 units of seed protein would convert to one unit of seed oil when increasing one unit of yield. Some reports suggest a weaker correlation. Based on 50-year average USA Uniform Trial data, Yaklich et al. (2002) noted a comparatively weak negative correlation between seed protein and seed oil (r = -0.39), and it was even weaker in the Northern USA Uniform Trial data (r = -0.33).

There are two main hypotheses proposed for this highly negative correlation of seed protein and seed oil, and neither one has yet to be convincingly ruled out. There is no clear, powerful evidence for discerning whether the inverse association of high protein with low oil (and *vice versa*) is due to (i) a single pleiotropic QTL with two alleles, wherein one allele simultaneously conditions high oil and low protein and another allele simultaneously conditions low oil and high protein, or (ii) two QTLs, wherein a high oil allele at an oil-controlling QTL is linked in repulsion phase with a low protein allele at a protein-controlling QTL (and *vice versa*). Many scientists have offered both hypotheses in their published papers; however, none have been able to critically reject one of these two hypotheses.

Most of the previous reports in the literature have documented a highly negative correlation between soybean seed protein and oil, which would seem to be difficult to overcome when breeding for higher contents of both traits; however, in some reported populations, the correlation was slightly weaker, suggesting some progress could be made towards improving seed oil without simultaneously incurring a substantive decrease in seed protein. The heritability of soybean oil content is high, which indicates that environmental effects do not have a huge effect on the phenotypic variance. If so, then soybean breeders might be able to use simple selection techniques with appropriate breeding methods to develop high oil breeding lines or cultivars. High seed oil cultivars would likely be of use to the biodiesel industry.

Molecular Markers in Plant Breeding

There are several different categories of markers that have been used to assist plant breeders in their crop genetic improvement programs. In the past, markers were limited to genes governing plant morphology or pigmentation. However, DNA markers have now become the marker of choice for breeders in their crop genetic studies, and have revolutionized the practical applications of plant biotechnology (Kumar et al., 2009). DNA markers are convenient because of their large numbers, distribution over the crop genome, and their naturally high polymorphism. To use DNA markers, one need only acquire small amount of tissue from any plant development stage.

Markers of the restriction fragment length polymorphism (RFLP) type were first used in soybean breeding and genetic studies in the late 1980s to construct genetic maps and to map genes controlling various traits (Keim et al., 1990; Dier et al., 1992). Since then, other types of DNA markers have been developed for the utilization in plant breeding programs. Random amplified polymorphic DNA (RAPD) markers and amplified fragment length polymorphism (AFLP) marker were also used for a period of time. However, simple sequence repeat (SSR) markers are now the DNA markers used by most soybean breeders in their breeding programs. Recently, single nucleotide polymorphism (SNP) markers have become available, and because of their huge numbers and the ability to genotype without using electrophoretic gel systems, soybean breeders and geneticists have moved quickly to use SNP markers in their work (Zhu et al., 2003; Choi et al., 2007). SNP markers are less informative than SSRs on a per locus basis, because SSRs can be multi-allelic. However, SNPs are much more abundant and genotyping hundreds of lines with thousands of SNP markers is now routine. Co-dominance is also important for information purposes. DNA marker types that exhibit co-

dominance, such as SNPs, SSRs and RFLPs (Brown and Caligari, 2008), are more useful than RAPDs and AFLPs, which are dominant markers.

A RFLP marker is generated by use of specific enzymes known as DNA Restriction Endonucleases. DNA is cleaved into fragments of variable lengths that can differ between two parents because of the presence and absence of DNA restriction enzyme sites. DNA size fractionation is achieved by gel electrophoresis. The parentally differently sized DNA fragment must be hybridized with radioactive labeled probes in order to visualize the DNA banding patterns on an electrophoretic gel. These probes are mainly species-specific, mono- or di-genic, and usually 0.5 – 3.0 kb in size (Kumar et al., 2009). The source of the probes is typically a cDNA library or a genomic library. RFLPs are almost invariably co-dominant markers, have high reproducibility, and are generally randomly distributed in the genome. However, species with larger genome need larger numbers of RFLP markers, which requires much more effort. This is not only time-consuming, but also labor-intensive, which makes RFLP analysis less desirable. However, RFLP markers were initially used in many early genetic studies due to their random distribution throughout the genome and abundant availability of different restriction enzymes (Neale and Williams, 1991). Indeed, RFLPs were first used to construct a genetic map in humans by Botstein et al. (1980). Keim et al. (1990) were the first authors to report a soybean RFLP-based genetic map.

SSR markers (also known as Microsatellites) have been extensively used in plant breeding programs and genetic studies. SSRs are based on tandem repeating sequences of a one to five base-pair repeat that are scattered throughout eukaryotic genomes (Powell et al., 1996). SSRs are frequently used to distinguish closely related genotypes because of

the multiple alleles that can be present at each of many SSR loci (Smith and Devey, 1994). Specific forward and reverse primers have to be designed for each flanking region of the repeat, which makes the initial development of SSR markers expensive; although primers designed for closely related species can sometimes be used for a given species. The advantages of SSR markers include co-dominance, high genomic abundance in eukaryotes, and a seemingly random distribution throughout the genome. Only small amount of DNA is required because genotyping is performed using a PCR-based reaction. High quality DNA also is not required due to the use of long PCR primers (generally 20-25 base-pair) and high reproducibility of SSRs. However, one of the main problems of SSR marker analysis is the need to develop of primers for any species that has not been sequenced. SSRs are considered as ideal markers in genetic mapping studies (Hearne et al., 1992). More than 1800 soybean SSR loci have been mapped to date (Cregan et al., 1999; Song et al., 2004; Hwang et al., 2009). In addition, SSRs have now replaced RFLPs as the soybean molecular of choice, because while SSR markers are mono-locus, each locus is potentially multi-allelic, which increases the opportunity for parental polymorphism compared to bi-allelic markers (Hyten et al., 2008). In the past two decades, SSR markers were routinely used to identify the locations of genomic segments with genes controlling major soybean agronomic traits.

Single nucleotide polymorphism (SNP) markers have now become more popular than SSRs for genotyping germplasm and segregating F_2 or RIL populations. SNPs have proven to be the greatest source of DNA polymorphisms in human beings. However, SNP marker development in plants has not been rapid as that in humans. The development of SNP markers is based on the fact that most genomic polymorphisms arise

from a single nucleotide change (i.e., point mutation). At the beginning of soybean SNP development, sequence variation searches for SNP markers was limited to specific genes or DNA fragments (Zhu et al., 1995). Zhu et al. (1995) screened a 400-bp fragment of RFLP probe in three different soybean genotypes and detected nine SNP loci. After this report, researchers began to actively search for SNPs (Coryell et al., 1999; Zhu et al., 2003; Van et al., 2004; and Hyten et al., 2006). Hyten et al. (2008) used the Illumina GoldenGate assay to illustrate the multiplexing of SNP allele identification (i.e., genotyping) at 96 to 1,536 soybean SNP loci in a single reaction over a 3-day period using genomic DNA from three RIL mapping populations: 'Minsoy' x 'Noir 1' and 'Minsoy' x 'Archer' from University of Utah, and 'Evans' x 'Peking' from University of Minnesota. The results showed that 342 SNP allelic data were obtained when 384 SNPs were evaluated (i.e., 89%). In 2010, Hyten et al. (2010) used the GoldenGate assay to map an additional 2,500 SNPs in the soybean genome. The authors also identified a set of 1,536 SNPs that were distributed more or less uniformly on all 20 soybean chromosomes.

To date, SSRs and SNPs have been the most extensively used markers for soybean breeding and genetic studies. Many research laboratories are still using SSRs for some genetic and breeding work because of the low cost. However, several USDA laboratories and laboratories from many universities are now devoted to SNP marker genotyping. Therefore, because of the high-throughput genotyping ability, SNP marker analysis will likely become the widespread technique for most plant genetic research and crop breeding programs.

Soybean Genetic Linkage Map

Soybean scientists have constructed genetic linkage maps by using molecular markers such as RFLPs, RAPDs, AFLPs, SSRs, and SNPs. The first soybean genetic linkage map was constructed by Keim et al (1990) using RFLP markers. A total of 150 RFLP markers were mapped to 26 genetic linkage groups based on an F_2 population derived from a cross between a *G. max* breeding line and a *G. soja* accession. Shoemaker and Olson (1993) later constructed a revised genetic linkage map of 25 linkage groups with 365 RFLPs, 11 RAPDs, three classical markers, and four isozymes loci based on an F_2 population derived from the same mating. Using the RFLP linkage map constructed in 1993, Shoemaker and Specht (1995) integrated several classical genetic markers and reported that half of the classical linkage groups could be associated with corresponding molecular linkage groups.

Cregan et al. (1999) was the first to construct a genetic linkage map that had 20 linkage groups which thereby corresponded with the 20 homologous pairs of soybean chromosomes. The authors mapped 606 SSRs, 689 RFLPs, 79 RAPDs, 11AFLPs, 10 isozymes, and 26 classical loci in one or more of three genotyped mapping populations, and assigned these loci to their respective linkage group. The three available mapping populations were developed by different research institutes: *G. max* x *G. soja* F_2 population by USDA/Iowa State University, 'Clark' x 'Harosoy' F_2 population by University of Nebraska, and 'Minsoy' x 'Noir 1' RIL population by University of Utah.

Subsequently, Song et al. (2004) reported an updated soybean integrated genetic linkage map. The authors used 420 new SSR markers, plus the 606 SSRs reported by

Cregan et al. (1999), in one or more of the five frequently used soybean mapping populations. These five mapping populations included the USDA/Iowa State University *G. max* x *G. soja* population 'A81-356-022' x 'PI 468916', University of Nebraska *G. max* x *G. max* population 'Clark' x 'Harosoy', and the three University of Utah populations 'Minsoy' x 'Noir 1', 'Mnsoy' x 'Archer', and 'Archer' x 'Noir 1'. The five maps were combined into one integrated genetic linkage map of 20 linkage groups using JoinMap software. To sum up, this linkage map spanning 2,523.6 cM of Kosambi map distance, consisted of 1,015 SSRs, 709 RFLPs, 73 RAPDs, 24 classical markers, six AFLPs, ten isozymes, and twelve other markers.

The soybean genetic linkage maps developed with RFLPs, AFLPs, RAPDs, and SSRs was expanded to include SNP markers in a soybean transcript map constructed by Choi et al. (2007), who were the first to report a soybean genetic linkage map using SNP markers. The authors mapped 1141 SNP loci in one or more of three RIL populations: the two University of Utah populations 'Minsoy' x 'Noir 1' and 'Minsoy' x 'Archer', and the University of Minnesota 'Evans' x 'PI 209332' population. SNPs were identified via the resequencing of sequence-tagged sites (STSs) which were developed from expressed sequence tags (ESTs). From a set of 9,459 PCR primers, 4,240 STSs were developed and were amplified and sequenced in each of six soybean genotypes: 'Archer', 'Minsoy', 'Noir 1', 'Evans', 'PI 209332', and 'Peking'. A total of 5,551 SNPs were identified (4,712 single base changes and 839 indels) in 2.44 Mbp of aligned sequence. 291 of 1,141 genes were mapped to positions within 72 of the 112 gaps of 5-10 cM in the SSR-based map reported by Song et al. (2004), while 111 of them were mapped to positions within 19 of 26 gaps of >10 cM. Adding these 1,141 SNP markers to the existing genetic

linkage maps benefited soybean breeders wishing to use SNPs in their QTL studies and in marker-assisted selection.

Recently, Hyten et al. (2010) reported the latest version of soybean integrated genetic linkage map (Consensus Map 4.0). These authors used 2,651 new SNP markers to genotype three mapping populations: the two University of Utah populations 'Minsoy' x 'Noir 1' and 'Minsoy' x 'Archer', and the University of Minnesota population 'Evans' x 'Peking'. There were a total of 5,500 genetic markers in this new genetic linkage map. The authors also selected a set of 1,536 SNPs distributed across the soybean genome to create a Universal Soy Linkage Panel (USLP 1.0) for high-throughput soybean QTL mapping. Compared to the genetic linkage map reported by Choi et al. (2007), which had 40 gaps of 5-10 cM and seven gaps of >10 cM, the new linkage map had only 18 gaps of 5-10 cM and one gap of >10 cM. Currently, the Soybean Consensus Linkage Map 4.0 is the standard marker genetic map used by researchers.

Quantitative Trait Loci (QTLs) Analysis

Soybean seed oil content is an important trait that is quantitatively inherited. Other traits such as yield, seed protein content, and disease resistance are also quantitatively inherited. If a trait is quantitatively inherited, the trait is likely to be governed by several genes that may vary in their effect on the phenotype (i.e., from major to minor). A fragment of DNA containing a gene governing a quantitative trait is known as quantitative trait locus (QTL). Since it is almost impossible to detect QTLs based solely on a phenotypic evaluation and existing morphological and pigmentation markers,

molecular markers such as SSRs and SNPs are used to locate the position of the QTLs on a linkage map (Collard et al., 2005).

To date, 73 QTLs have been detected for seed oil content (SoyBase, 2011). Diers et al. (1992) developed a population from a cross between a *G. max* experimental line A81-356022 and a *G. soja* accession PI468916, and this population was used to generate RFLP map. In this population, two major oil QTLs were detected on LG-I (Chr 20) and on LG-E (Chr 15). Homozygous lines for *G. max* allele at the most significant marker linked to the QTL had seed oil contents that were greater by 17 g kg-1 for LG-I (Chr 20) QTL, and by 11 g kg-1 for LG-E (Chr15) QTL, than homozygous lines for the *G. soja* allele at that same marker. The results showed that the *G. max* alleles at significant oil loci were associated with greater oil content compared to *G. soja* alleles; this was expected because these two parents were significantly different for seed oil content, and the *G. max* parent had a much higher oil content.

Mansur et al. (1993a) detected two RFLP-flanked marker intervals, T153-A111 and BCI-A315, each containing an oil QTL in an $F_{2.5}$ population from 'Minsoy' x' Noir 1'. These two intervals mapped to LG-A2 (Chr 8) and LG-L9 (now LG-K; Chr 9), respectively. The T153-A111 interval explained 36% of phenotypic variance, and BCI-A315 interval explained 24% of phenotypic variance. The 'Minsoy' allele at the two seed oil loci had decreased the seed oil. There also was an unlinked RFLP locus, K1, now known to map to LG-C1 (Chr 4), that was associated with seed oil content in a 'Minsoy' x 'Noir 1' derived $F_{2.5}$ population. Lark et al. (1994) also detected a combined protein and oil QTL linked to a RFLP locus R183_1 in RIL derived from 'Minsoy' x 'Noir 1'. Marker R183_1 is located on LG-A1 (Chr 5).

Lee et al. (1996) mapped seed oil QTLs in two populations. In the population of 120 F_4-derived lines from a cross of 'Young' x 'PI 416937', six seed oil QTLs were reported on LGs E (Chr 15), J (Chr 16), L (Chr 19), and R (now LG-D2, Chr 17), which three of the six oil QTLs were detected on LG-R (now LG-D2, Chr 17) . In the population of 111 F2-derived lines of 'PI 97100' x 'Coker 237', five seed oil QTLs were detected on LGs C1 (Chr 4), G (Chr 18), and H (Chr 12), which three of the five oil QTLs were detected on LG-G (Chr 18). At the locus L154-2 on LG-G (Chr 18) in the 'PI 97100' x 'Coker 237' population, heterozygotes had a higher seed oil percentage than homozygotes, suggesting overdominance at this QTL. None of seed oil QTLs was common in both populations suggesting that all 11 seed oil QTLs were population-specific. The markers linked to seed oil QTLs on the LG-E (Chr 15) were close to the markers linked to the seed oil QTLs reported by Diers et al (1992. The A069_2 locus on LG-E (Chr 15) associated with a seed oil QTL in 'Young' x 'PI 416937' was near the A374_1 locus and 5cM from A203_1 locus associated with seed oil QTL in 'A81-356002' x 'PI 468916'.

Brummer et al. (1997) developed eight different populations of F_2-derived lines and evaluated these for protein and oil at various test locations. They detected RFLP markers associated what the authors called "environmentally stable" QTLs for soybean oil content in seven linkage groups: LGs A1 (Chr 5), A2 (Chr 8), B1 (Chr 11), C2 (Chr 6), G (Chr 18), H (Chr 12), and K (Chr 9). In this study, an "environmentally stable" QTL was defined as a QTL detected in at least two of the three individual test years, and also in the 3-year average data, using a Type I error criterion probability of $P \leq 0.05$. The authors also detected ten "environmentally sensitive" oil QTLs in six linkage groups.

These QTLs were not considered as useful as the "environmentally stable" QTLs, because specific environment was presumed to be required for the expression of the "favorable" allele. Relative to all QTLs, one marker, A584_1, was detected in more than one population, while other markers were only detected in one population suggesting that most of "environmentally sensitive" QTLs also were population-specific. Although the seed oil QTLs detected in this study and those reported by Diers et al. (1992) did not match with respect to linkage group map positions, several oil QTLs mapped on LG-A1 and LG-E in this study were close to fatty acid QTL detected by Diers and Shoemaker (1992). The RFLP marker T153-1 on LG-A2 (Chr 8) for seed oil was also detected by Mansur et al. (1993).

Qiu et al. (1999) identified one RFLP marker, B072 on LG-H (Chr 12), associated with seed oil QTL in a $F_{2.3}$ population derived from 'Peking' x 'Essex'. This QTL only explained 21% of phenotypic variance, suggesting that there might be other undetected QTLs controlling seed oil content. The authors reported that RFLP marker B072 was also associated with seed protein content. The authors reported that this favorable allele for protein and oil was from 'Essex' parent, suggesting that 'Essex' could be a potential breeding source for both high seed protein content and seed oil content. In this study, however, soybean seed protein and oil had a correlation coefficient r = -0.886, indicating the usual inverse correlation between these two traits, casting doubt on the QTL parental A and B assignment made by the authors. More research is needed to confirm this QTL, but if the 'Essex' allele does improve both protein and oil, then this QTL could be useful for solving the problem of inverse correlation between seed protein and oil content.

Orf et al. (1999) developed three RIL populations from the crosses of 'Minsoy' x 'Noir 1' (i.e., MN population), 'Archer' x 'Minsoy' (i.e., MA population), and 'Noir 1' x 'Archer' (i.e., NA population) to identify QTLs for various agronomic traits. A RFLP marker T155_1 on LG-A1 (Chr 5) was linked to seed oil QTL in MN population. This marker was also reported earlier by Mansur et al. (1996).The SSR marker SOYGPATR on LG-A1 (Chr 5) was linked to seed oil QTL in the MA population. However, these two markers were also linked to seed protein QTLs in this study. There were three other markers linked to seed oil QTLs: SSR Satt174 on LG-A1 (Chr 5) in the MA population, SSR Satt432 on LG-C2 (Chr 6) in the NA population, and RFLP A489_1 on LG-L (Chr 19) in the NA population.

Specht et al. (2001) used an RIL population derived from a cross of 'Minsoy' x 'Noir 1' to identify QTLs for drought tolerance and seed protein/oil content. They detected oil QTLs that had been identified and reported in prior studies. The study indicated that increasing water stress tended to increase oil content. However, they did not detect any seed oil *beta* QTLs that might control the change in seed oil that occurred with variable amounts of seasonal crop evapotranspiration. Such QTLs would be useful for reducing unwanted changes in seed oil induced by variation in the timing of irrigation applications during the growing season.

Csanádi et al. (2001) identified seed oil QTLs in a cross of the two early maturing soybean cultivars 'Ma. Belle' and 'Proto'. The authors constructed a linkage map with a marker set of 113 SSRs, six RAPDs, and one RFLP that were segregating in the 82 individuals of the F_2 population. Three seed oil QTLs were detected. The SSR markers Satt020 on LG-B2 (Chr 14), Satt196 on LG-K (Chr 9), and Satt562 on LG-I (Chr 20)

were detected to be closely linked to seed oil QTLs. SSR Satt196 was linked to both seed oil QTL and seed protein QTL. The authors also reported a close linkage of seed oil and seed protein for marker Sct_028 on LG-C2 (Chr 6). The negative correlation between oil and protein content in this study was typical of that reported by others (Lark et al., 1994; Lee et al., 1996; Sebolt et al., 2000).

Chapman et al. (2003) detected soybean QTLs for agronomic traits in an F_2 population and an $F_{4.6}$ population of 177 lines derived from a cross of 'Essex' x 'Williams'. The authors identified two QTLs for seed oil content in the F_2 population, one linked to SSR Satt251 on LG-B1 (Chr 11) and the other one linked to SSR Satt14 on LG-D2 (Chr 17). Satt251 was also linked to a protein QTL in the F2 population. No oil QTLs were detected in the $F_{4.6}$ population.

Chung et al. (2003) identified a high oil allele from PI 437088A, a *G. max* accession at a QTL located on LG-I (Chr 20). The QTL for oil mapped to an interval flanked by SSR markers Satt496 and Satt239. The closest marker linked to oil QTL was OPAW13a, and it was just two cM from Satt239. This high oil allele unfortunately was also a low protein allele. The strongest two RFLP markers associated with seed oil detected by Diers et al. (1992), K011 and A407, mapped 1.8 and 1.7 cM above Satt239, respectively. Sebolt et al. (2000) reported that RFLP marker A144, which mapped 7.1 cM above Satt239, was associated with seed oil.

Hyten et al. (2004) reported seed oil QTLs were detected in a 131 F_6-derived RIL population created from a cross of 'Essex' and 'Williams', when the RILs were evaluated in six different testing environments. The authors genotyped the RILs at 100 SSR marker

loci. They identified six oil QTLs in this population: one on LG-C2 (Chr 6), one on LG-D1a (Chr 1), one on LG-D2 (Chr 17), two on LG-L (Chr 19), and one on LG-M (Chr 7).

Fasoula et al. (2004) followed up a prior report (Lee et al., 1996) on seed oil QTLs in two soybean populations with an experiment aimed at confirming those prior detected QTLs. The two populations created by Lee et al. (1996), 'Young' x 'PI 416937' and 'PI 97100' x 'Coker 237', were created again by new matings. In the new 'PI 97100' x 'Coker 237' population, two of the three QTLs reported by Lee et al. (1996) were confirmed. These were the two QTLs linked to RFLP markers A063-1 on LG-C1 (Chr 4) and A566-2 on LG-H (Chr 12). In the new 'Young' x 'PI 416937' population, only one of three oil QTLs reported previously was confirmed. Two SSR markers, Satt398 and Satt313, were linked to RFLP marker A023-1 that was linked to the oil QTL on LG-L (Chr 19). The authors indicated that those unconfirmed QTLs were possibly false positives (Type I errors) in the original populations. These results confirmed the necessity of validating QTLs in re-created populations before utilizing the QTL alleles in future plant improvement programs.

Kabelka et al. (2004) detected QTLs for increasing seed yield or seed oil in North American cultivars. The authors developed a population of 167 F_5-derived lines from the mating of 'BSR 101' x 'LG82-8379' and divided the population into three sets based on the RIL maturities. They detected three seed oil QTLs on LGs C1 (Chr 4), D1b (Chr 2), and J (Chr 16). Two of three QTLs, linked to markers Satt338 and Satt157, explained 10% of the phenotypic variation for seed oil concentration in combined analysis. One QTL on LG-J (Chr 16) was detected only in one set and explained 16% of phenotypic

variation. A total of 26% of phenotypic variation was explained by the three QTLs in Set1.

Panthee et al. (2005) reported the seed oil QTLs detected from a population of 101 F_6-derived recombinant inbred lines (RILs) created from the mating of 'N87-984-16' x 'TN93-99'. Only 94 SSR markers were polymorphic amongst the 585 SSR markers assayed. Three allegedly new seed oil QTLs were detected based on the linkages to markers Satt274 on LG-D1b (Chr 2), Satt420 on LG-O (Chr 10), and Satt479 on LG-O (Chr 10). A seed oil QTL linked to marker Satt317 on LG-H (Chr 12) had been reported in prior studies. Although many previous studies have reported the detection of QTLs within a given linkage group, the map position of the previous reported markers were far away from the three markers reported in this paper, suggesting that these three QTLs were newly discovered. However, marker Satt317 was only 20 cM away from the RFLP marker B072_1 reported by Qiu et al. (1999). These four QTLs explained 11.8% (Satt274), 9.4% (Satt317), 15.0% (Satt420), and 12.0% (Satt479) of phenotypic variation, respectively.

Panthee et al. (2006) detected several QTLs for fatty acid composition in soybean oil from a population of 101 F_6-derived recombinant inbred lines (RIL) from a cross of 'N87-984-16' x 'TN93-99'. A total of seven markers associated with fatty acid composition were detected in this study: Satt133 on LG-A2 (Chr 8) and Satt 537 on LG-D1b (Chr 2) associated with palmitic acid; Satt168 on LG-B2 (Chr 14) and Satt249 on LG-J (Chr 16) associated with stearic acid; Satt263 on LG-E (Chr 15) associated with oleic acid and linolenic acid; Satt185 on LG-E (Chr 15) associated with linoleic acid; and Satt235 on LG-G (Chr 18) associated with linolenic acid. The phenotypic variation an

individual QTL explained ranged 10-22.5%. There were not many markers associated with multiple forms of fatty acids, indicating that any improvement in soybean oil quality would likely require enhancing each fatty acid with independent markers.

Monteros et al. (2008) reported QTLs associated with oleic acid content in a population of 259 $F_{2.3}$ from a mating of 'G99-G725' x 'N00-3350'. All the QTLs detected in this population were confirmed in a $F_{2.3}$ population of 231 lines form a cross of 'G99-G3438' x 'N00-3350'. There were six QTLs associated with oleic acid, and these were located on LGs A1 (Chr 5), D2 (Chr 6), G (Chr 18), and L (Chr 19). All of the six QTLs were confirmed in 'G99-G3438' x 'N00-3350' population. The phenotypic variation explained by each individual QTL ranged from 4.0 to 25%.

The results reported in the papers listed above show that many oil QTLs have been detected though only a few have been confirmed to date. Recently, many researchers have focused on particular fatty acid composition in soybean seed oil, searching for QTLs that would not only improve total seed oil content but also increase specific desirable fatty acid levels. Finding QTLs for both total oil content and fatty acid composition will benefit the future researchers using marker-assisted selection (MAS) in soybean oil breeding programs.

Selective Genotyping

Selective genotyping (SG) is a method developed for the detection of QTLs with minimal genotyping, so as to reduce the cost in time, and the expense of the latter. SG requires a large number of progeny or RILs to be phenotyped to ensure a reasonable

statistical power for QTL detection. In any SG study, the only genotyped individuals are those with extreme phenotypes (Lebowitz et al., 1987; Lander and Botstein, 1989; Darvasi and Soller, 1992; Darvasi and Soller, 1994). Generally, only the individuals from the highest and the lowest fraction of the phenotypic trait distribution in the population are ultimately genotyped (Darvasi, 1997). As a matter of fact, as Lander and Botstein (1989) noted, the most informative individuals in any population are those whose genotype is inferred by their phonotype, which basically defines those individuals with phenotypes that deviate the greatest from the phenotypic mean. One of the main limitations of the SG approach is that only one trait can be analyzed at a time, because if many uncorrelated traits are analyzed simultaneously, most of the individuals from the population will have to be selected for SG, and there will be no reduction in genotyping numbers (Darvasi, 1997). However, if two highly-correlated traits are analyzed at the same time, the selected extreme individuals for each trait are likely to overlap, making SG possible for both traits.

There have been many researchers reporting different SG portions based on various experimental populations. Lander and Botstein (1989) suggested that the highest 17% and the lowest 17% of individuals (i.e., individuals with phenotypes exceeding a plus or minus one standard deviation from the population mean), amounting to 34% of the population would account for 81% of the total linkage information. The authors also mentioned that the number of individuals needed to be genotyped decreases when the phenotypic difference between the two parental lines increases (i.e., since the F_2 phenotypic variance also would increase). Darvasi and Soller (1992) reported that to maximize SG efficiency, it may not be effective to select more than 50% of the total

population (i.e., the highest 25% and the lowest 25%). Ayoub and Mather (2002) suggested that a selection of 10% or 20% of the total population (i.e., 5% or 10% of each tail) was sufficient to detect all of the QTLs (with a SG technique), that had been detected previously by an interval mapping technique applied to the same population that had been completely genotyped.

Lebowitz et al. (1987) suggested several theoretical concepts for predicting the difference in marker allele frequencies between the lowest and the highest tails of an F_2 population. They provided the following equation:

$$\delta_M = \frac{(i_p)(2a)(m1)(m2)}{\sigma_p}$$

where,

i_p = the standard selection differential for the decile selection in an F_2 population (i.e., i_p of 10% selection = 1.755),

a = the additive effect of the parental alleles segregating at the QTL,

m1 = m2 = 0.5, the population frequencies expected for the two parental alleles at a given locus for an F_2 population,

σ_p = the population's phenotypic standard deviation.

The foregoing equation is only for a small QTL effect approximation. It can be technically improved by dividing the equation by the following:

$$1 - [\frac{(i_{p)(a)}}{\sigma_p}]$$

The standard error ($SE_{\delta M}$) of an observed change in marker allele frequency between the two tails is:

$$SE_{\delta M} = \sqrt{\{\frac{[(2)(m1)(m2)]}{[(2)(n)]}\}}$$

where,

n = number of tail marker alleles,

m1 = m2 = 0.5, the population frequencies of the two parental alleles at a given locus for the F_2 population.

The statistical power of selective genotyping is related to phenotypic standard deviation and the additive effect of the QTL detected in an F_2 population. When using selective genotyping as an approach to detect QTL for a given trait in a given population, one desires to know the appropriate number of F_2 individuals to be phenotyped and the fraction of the extreme progeny that should be genotyped for a given desired power for the detection of a QTL of some specified additive effect. To calculate the power of selective genotyping for detecting linkage between QTL and marker the following equation can be used:

$$Z_\beta = \left[\frac{\delta_{a(F2)}}{SE\delta_{A(F2)}}\right] - Z_\alpha$$

where,

Z_α = the ordinate of a normal curve corresponding to the likelihood of the chosen level of α error (i.e., Type I error),

Z_β = the ordinate of a normal curve corresponding to the likelihood of the desired level of β error (i.e., Type II error).

The power of the QTL detection is calculated by: Power = 1- Z_β.

There are no formal standards for choosing a power value; however, most researchers would choose population parameters that would provide a power value of 0.8, which would be equivalent to a type II error probability of $\beta = 0.20$ (assuming a type I error probability of $\alpha = 0.05$).

THESIS RESEARCH OBJECTIVES

In this thesis research project, three $F_{2.4}$ populations segregating for seed oil content (up to 550 individuals each) were created: two were created by the matings of two high oil breeding lines (RMLPC1-311-128-128 and U06-103459) with the normal seed oil cultivar (Williams 82), and the other one was derived from the mating of two high oil lines with each other.

My thesis research objectives were:

(1) Measure the seed oil content for each $F_{2.3}$ progeny of the three populations by two replications of Near-infrared spectroscopy (NIR) on 30 seed samples.
(2) Measure the seed oil content for each $F_{2.4}$ progeny in order to confirm the heritability of seed oil content.
(3) Use the selective genotyping method by phenotyping about 450 to 500 progeny in each population, but genotyping only the extreme quintiles with1536 SNP markers, to determine if one or more seed oil QTLs could be detected and if any of these QTLs had little or no pleiotropic impact on seed protein content.

MATERIALS AND METHODS

Parental Germplasm

Plant materials were selected based on seed yield and seed oil content performance. The two high seed oil lines used here were obtained from Dr. George Graef (Department of Agronomy, University of Nebraska-Lincoln).

The soybean line RMLPC1-311-128-128 is a high-oil Maturity Group III breeding line from the Cycle 1 of the Random Mating Low Protein (RMLP) Population developed in Dr. Graef's University of Nebraska soybean breeding program. The RMLP population was created by using *ms2* male sterility for intermating among high-protein lines. In a trial conducted by Dr. Graef, the soybean line RMLPC1-311-128-128 yielded 3,925 kg/ha (i.e., 62.8 bu/ac) and had an average seed weight of 14.7g/100 seed (i.e., 3,086 seeds/lb). The seed protein content was estimated to be about 361 g kg^{-1} (i.e., 36.1%) and the oil content was about 248 g kg^{-1} (i.e., 24.8%).

The U06-103459 breeding line is a high oil Maturity Group II line developed from high-yield and high-oil matings. U06-103459 was developed from a 2004 High Oil mating of the parent NE2801 (a high yielding cultivar release) with U01-290680 (a high oil breeding line). NE2801 was derived from an intermated population using *ms2* male sterility to facilitate intermating. U06-103459 is a late Maturity Group (MG) II cultivar and in a trial conducted by Dr. Graef, it yielded 3,830 kg/ha (i.e., 60.9 bu/ac). U01-290680 was derived from the high-oil mating of NE3001 (high-yielding released cultivar) with HOL-833. NE3001 is a MG III cultivar that in a trial conducted by Dr. Graef,

yielded 3,893 kg/ha (i.e., 61.9 bu/ac). The seed protein content was estimated to be about 388 g kg^{-1} (i.e., 38.8%) and the oil content was about 250 g kg^{-1} (i.e., 25.0%).

Williams 82 is a *Glycine Max* accession developed in Illinois and released in 1982 as an improvement of the older Williams cultivar released in 1971 (Bernard et al., 1988). Information about Williams 82 is available from the Germplasm Resources Information Network (GRIN) website (http://www.ars-grin.gov/cgi-bin/npgs/acc/display.pl?1413607). Information on the three parents in provided in Table 1.

Population Development

Matings

During the spring of 2008, 50 seeds of each parent were planted into a 2.5-m row of a crossing block on UNL East Campus. There was a 60-cm spacing between rows in order to allow working space for pollinations. When a female parent flowered, it was mated to the synchronously flowering male parent. The first crosses were attemped in 2008, but F_1 seeds were not obtained. Seeds of the three parents were then sent to the Puerto Rico winter nursery to re-try the matings. In December 2008, 50 seeds of each parent were planted into a 2.7-m row in Puerto Rico. There was a 0.9-m spacing between rows in order to allow working space for pollinations. Three planting dates were used to provide some overlap in flowering duration among the three parents. Three matings (U06-103459 x RMLPC1-311-128-128, Williams 82 x RMLPC1-311-128-128, and U06-103459 x Williams 82) were made, as were reciprocal matings. About 20 pollinations were attempted for each of the three matings, and putative F_1 seeds were successfully

obtained. These were code-named “UX2427” (U06-103459 x RMLPC1-311-128-128), “UX2428” (Williams 82 x RMLPC1-311-128-128), and “UX2430” (U06-103459 x Williams 82). For the reciprocal matings, a letter “B” was added after the population name (e.g., UX2427B is the reciprocal cross of UX2427 mating) to distinguish the putative reciprocal F_1 seeds from the putative forward cross F_1 seeds. For the UX2427 matings, five pods from forward crossing and one pod from reciprocal crossing were obtained. For the UX2428 matings, four pods from forward crossing and three pods from reciprocal crossing were obtained. For the UX2430 matings, six pods from forward crossing and two pods from reciprocal crossing were obtained. The F_1 seeds in the foregoing pods obtained from each mating were placed into packets labeled by cross, pod number, and seed number.

F_1 Generation

During the summer of 2009, the F_1 seeds from each mating, plus seeds of the parents of the three matings were grown in separate rows in the UNL East Campus crossing block. A trifoliolate leaf from each emerged putative F_1 plant was collected for subsequent DNA extraction. An F_1 hybridity test was conducted on the DNA using parentally polymorphic SSR markers (Table 2). Only those F_1 plants confirmed by the SSR analysis to be true hybrids were collected to be individually threshed to obtain the F_2 seed (i.e., $F_{1.2}$ progeny).

F_2 Generation

F_2 seeds produced from each confirmed F_1 plant were planted in the UNL East Campus greenhouse in the winter of 2009-2010. This greenhouse planting involved 524

F_2 seeds (263 seeds from the eight F_1 plants of the forward mating and 261 seeds from the one F_1 plant of the reciprocal mating) obtained from the nine total UX2427 F_1 plants, 485 F_2 seeds (230 seeds from the five F_1 plants from the forward mating and 255 seeds from the eight F_1 plants of the reciprocal mating) obtained from the 13 total UX2428 F_1 plants, and 508 F_2 seeds (258 seeds from the ten F_1 plants of the forward mating and 250 seeds from the two F_1 plants from the reciprocal mating) obtained from the 12 total UX2430 F_1 plants (Table 2). Six F_2 seeds were planted into each 28-cm diameter by 28-cm deep pot filled with a 1:1 mixture of steam-sterilized soil and Metro-Mix 360 soil-free media. Five seeds were planted at the "five-point star" positions near the circumferential rim of a pot, with a sixth seed planted at the center of the "star". A total of 30 seeds of each parent were also planted in the greenhouse (five seeds per pot; six pots per parent). An automated drip-irrigation system was used to supply water as needed for each pot. When the second trifoliolate leaves of most of the F_2 plants had reached a fully expanded stage, a numbered tag was wired to each F_2 plant between the first and second trifoliolate nodes for F_2 plant identity purposes (i.e., labeled with mating code, and F_1 and F_2 plant numbers). Thirty parental plants were also tagged with numbered tag. Mature F_2 plants were individually threshed and their $F_{2.3}$ seeds were individually packeted with care to ensure that the numbered packet label matched the numbered tag on each plant. Parental seeds were also harvested and packeted individually for use as reference phenotypes during the later seed protein and oil phenotyping analyses.

F_3 Generation

From each $F_{2.3}$ seed harvest packet, 30 seeds were selected and packeted into a 2010 spring planting packet, but if there were less than 30 seeds, all harvested seed was

placed in the planting packet. The planting packet was labeled with a barcode indicating the mating, F_1 plant, and F_2 plant number associated with F_3 seeds inside. A planting map was prepared that had 48 rows by 48 tiers. The top three and bottom three tiers were border rows planted with Nebraska cultivar NE3001. The leftmost three rows and rightmost three rows were also border rows planted with NE3001. The central 42 rows by 42 tiers were then divided into 98 blocks, with each block consisting of three tiers of six rows (i.e., a total of 18 rows). Each of the two parents of a given population was randomly assigned to one row in each given block, with the 16 $F_{2.3}$ progenies of a given population randomly assigned to the remaining rows in the block. Each population was assigned to 31 to 33 blocks depending on the number of individual rows. This $F_{2.3}$ progeny row experiment with two parent rows per block is an augmented design commonly used by plant breeders. The experiment was planted in the East Campus Field M in the spring of 2010. Before harvesting, one plant from each progeny row was tagged with a label containing a barcode indicating the field block location of the progeny row and progeny ID information. Parent plants from each block were also labeled with tier/row number to provide information of their location in the blocks. Progeny rows (and parental rows) were collected and threshed on a per row basis to obtain $F_{2.4}$ seed. Figure 1 shows the complete development of three F_2-derived F_4 populations.

SSR Marker Analysis – F_1 Plant Tissue

SSR markers were used for F_1 hybridity confirmation in this study (Table 2). These SSRs were discovered, mapped, and reported by Cregan et al. (1999). A set of

eight SSR markers was selected based on primer availability, probable parental polymorphism, and possible linkage to strong oil QTLs that have been previously mapped. According to the data from SoyBase, several linkage groups (LGs) possibly contain a strong oil QTL, such as LGs A1 (Chr 5), C1 (Chr 4), E (Chr 15), G (Chr 18), H (Chr 12), I (Chr 20), and L (Chr 19). The three parents were initially screened with this set of SSR markers to identify which SSRs were parentally polymorphic in each mating. The final set of selected markers, which were polymorphic between parents, were used for F_1 hybridity confirmation (Table 2).

The SSR analysis was conducted using PCR amplification. The PCR reaction mix consisted of 50ng of genomic DNA, 0.2µM primer each of the paired forward and reverse primers, 1.5 mM $MgCl_2$, 5X of reaction buffer (50mM KCl, 10mM Tris-HCl, 0.1% triton X-100), 0.7 units of DNA Taq polymerase, and 0.15 mM of each of the four dNTPs. The reaction mix was pipetted into a 96-well reaction plate, and then the plate was sealed with a polypropylene-based film in order to prevent evaporation. The PTC-100 Programmable PCR thermocycler (MJ Research, Watertown, MA) was used to accomplish PCR reaction. The PCR schedule consisted of 31 cycles of a 3-step thermocycler reaction. The three steps were: (i) 94℃ for denaturation for 25 s; (ii) 47℃ for annealing for 30 s; and (iii) 68℃ for extension for 25 s, with the last cycle at 68℃ followed by an incubation at 4℃.

A 2.5% (W/V) agarose (Amresco, Solon, OH) gel was prepared for each population (i.e. total three gels for this oil QTL study). The PCR products were then loaded on the gel. A 0.5X TBE solution served as a running buffer. The gel was run at a constant 70V for 5 hours. The gel was stained in ethidium bromide for 15 minutes and

then de-stained in distilled, deionized water for 15 minutes. The gels were exposed under UV light and the banding patterns were captured by an image analysis system (GelDoc2000, BioRad, Hercules, CA) and printed on thermal-sensitive photography paper. Parental and F_1 plant banding-patterns were scored "A" (homozygotes of the forward cross male parent), "B" (homozygotes of the forward cross female parent), and "H" (heterozygote of F1) for each primer locus.

Phenotypic Trait Evaluation, Measurement, and Analysis

$F_{2.3}$ Seed

During the spring of 2010, the F_2 plant-derived F_3 seeds (i.e., $F_{2.3}$ seed progenies) were evaluated for seed oil and protein content using the near-infrared reflectance (NIR) analyzer (Infratec 1241 Grain Analyzer) located in the Stewart Seed Laboratory. The oil, protein, moisture, and fiber content were evaluated simultaneously. Seed oil and protein content were measured by reflectance of electromagnetic radiation in the near infrared region of the spectrum and on a 13% (130 g/kg seed) seed moisture basis. The cuvette with two transparent glasses on the opposite sides is normally filled entirely with seeds (about 150 seeds are needed to fill up the whole cuvette, but this depends on the seed diameter), and then 10 sub-sample assay readings are obtained with each cuvette sample. In this study, however, most of the $F_{2.3}$ progenies had fewer than 150 seeds, so only 30 seeds from each $F_{2.3}$ envelope were poured into cuvette (all seeds were used for evaluation for those envelops with less than 30 seeds). Only three sub-sample reading were carried out because of the deficiency of F_3 seed. Parental seeds from 2009-2010

winter greenhouses were also NIR-evaluated. Parent line RMLPC1-311-128-128 did not reproduce well when grown in the greenhouse with metal halide lamps, so its plant progenies were bulked to get at least 30 seeds per cuvette sample.

The three populations were evaluated separately in time, and the two parents of the given population were also evaluated with the progenies at the same time as the reference phenotypes. For $F_{2.3}$ progenies, two replications of NIR analysis were conducted to assess measurement accuracy. Six packets of check samples were prepared: 30 and 100 seeds of RMLPC1-311-128-128, 30 and 100 seeds of U06-103459, and 30 and 100 seeds of Williams 82. Checks were used every hour to ensure that the NIR analyzer operated within performance standards during the time it was used.

$F_{2.4}$ Seed

Because of the lack of sufficient F_3 seed in from the $F_{2.3}$ progenies, a generation advance to $F_{2.4}$ progenies was conducted in the summer of 2010 in order to generate greater amounts of seed for use in a subsequent NIR analysis. For the seed oil and protein phenotyping of the $F_{2.4}$ seeds harvested in the fall of 2010, a one-replicate of NIR analysis was performed on a random sample of seed harvested from each $F_{2.4}$ progeny row of the three populations. The $F_{2.4}$ seed numbers were sufficient to fill the NIR cuvette with seeds. Ten sub-sample readings were performed on nearly all of the $F_{2.4}$ progeny lines (the few $F_{2.4}$ progeny packets with less than a full-cuvette seed amount were marked as such on the NIR result files). Parental seed samples from the same field test were also NIR-evaluated. The NIR analysis of $F_{2.4}$ progenies was accomplished in the following manner: The $F_{2.4}$ row seed bags were arranged in field block design order in the laboratory to allow the

seed oil/protein content of given *each block of entries to be NIR-assayed within a contiguous block of time*. This procedure ensured that any laboratory or NIR instrument environmental effects (e.g., seed moisture and seed temperature differences) were confounded with the field block differences, for analysis of variance purposes. Seven packets of seed sample checks (low seed oil lines: B1112 and B1027, normal seed oil lines: Williams 82 and NE3001, and high seed oil lines: RMLPC5S2-2006-56, RMHOC5S2-41-28-15, and U06-103459-31) were analyzed three times a day (8am, 12pm, and 5pm) to monitor the within-day and between-day repeatability of the NIR analyzer. To assess the precision (i.e., repeatability) of NIR measurements on a larger sample than the checks, the UX2428 population packets were re-assayed, so that the second-replication data values could be compared with the first-replication data values by linear regression analysis.

Leaf Collection and DNA Extraction Procedures

In the summer of 2009, one trifoliolate leaf was collected from each F_1 plant and stored in a 96-well plate, with special care to match the plate well number and tagged F_1 plant. Leaf tissues were stored at -20℃ until subsequent DNA extraction and SSR analysis.

During the winter of 2009-10, three leaflets were collected from each of the F_2 plants of each of three populations grown in the UNL greenhouse and stored in 96-well plates. One of the three leaflets was used for subsequent DNA extraction, whereas the other two were retained as backup. The three parents were planted in the greenhouse

during the winter of 2010-11 in order to get the fresh leaflets of those as well. All leaf tissue was stored at -80°C until subsequent DNA extraction. After the phenotyping of the $F_{2.4}$ seed progenies was completed, the progenies in each population were ranked by their oil content. Because the project was designed with reciprocal matings, the ranking was conducted within the $F_{2.4}$ progenies traceable to the forward mating, and then again within the $F_{2.4}$ progenies traceable to the reciprocal mating. The highest and lowest deciles were selected from both forward cross and reciprocal cross in the following manner. Because of the limited space in the 96-well plate, the 23 highest and the 23 lowest $F_{2.4}$ progenies were selected relative to seed oil content amongst those tracing to the forward mating as were a similar set of the 22 highest and the 22 lowest amongst those tracing to the reciprocal mating. As a result, there were 45 high oil selections and 45 low oil selections from each population of more than 450 progenies. Although a high/low decile selection genotyping had been originally planned, the actual percentage of the progenies chosen for selective genotyping in each population was somewhat less than 10%, specifically 45/507 = 8.88% for UX2427, 45/473 = 9.51% for UX2428, and 45/483 = 9.32% for UX2430. Leaf tissue of the chosen 90 progenies (45 high and 45 low) in each population had to be transferred from the leaf collection plates to 90 wells on one 96-well plate per population. Of the six remaining wells, two were used for leaf tissue of each of the two parents and two were reserved for the leaf tissue collected from an F_1 plant from the forward cross (F1A) and from an F_1 plant from the reciprocal cross (F1B). Replication of parents and F_1 plants is necessary in case of genotyping failure.

The leaf tissue extraction procedure employed BioSprint 96 DNA extraction kits that use the MagAttract magnetic-particle technology for DNA purification. The DNA

binds to the silica surface of MagAttract magnetic particles. The DNA bound to the magnetic particles is then washed with alcohol-containing buffers or ethanol. Then, the Tween wash improves the purity of the DNA. Finally, the purified DNA is dissolved in TE buffer for storage.

For the actual DNA extraction using the BioSprint instrument, six 96-well plates were necessary for each population. Plate 1 contained 200µl of lysate, 200µl of isopropanol, and 20µl of MagAttract Suspension G. Plate 2 contained 500µl of buffer RPW (with RNase and isopropanol). Plate 3 and Plate 4 contained 500µl of ethanol. Plate 5 contained 500µl of the wash with 0.2% Tween, which was for purpose of purifying DNA. Finally, Plate 6 contained 200 µl of TE, which was for the purpose of dissolving purified DNA.

Before the DNA extraction, a lysate had to be prepared. For the lysate preparation, 3 to 6 beads and 400µl of RLT were added to each well of three 96-well plates with leaf tissue transferred from the original collection plates. The plates were shaken for 5 minutes, and then centrifuged at 4100 RPM for 5 minutes. 200µl of the liquid residing above the leaf tissue was transferred to Plate 1 before adding 200µl of Isopropanol and 20µl of MagAttract Suspension G.

The six foregoing plates along with a collection microtube were placed in chronological sequence order in the BioSprint 96, and the BioSprint 96 was then turned on for automated DNA extraction. The purified DNA was finally dissolved in 200µl of TE (Plate 6), and immediately stored in 4°C.

Electrophoresis Protocols

Gel electrophoresis was used for qualification and quantification of genomic DNA *before* it was used for the SNP analysis. A 5µl aliquot of each liquid genomic DNA sample (including a sample of each of the two parents, plus a sample of the F1A and F1B hybrids) was transferred to a 96-well reaction plate. A 2.5µl aliquot of 5X reaction buffer (50mM KCl, 10mM Tris-HCl, 0.1%triton X-100) was also loaded into the 96-well reaction plate. A 1.0% (W/V) agarose gel (Amresco, Solon, OH) was prepared for each population. The 7.5µl DNA-buffer sample mix was loaded into a gel and electrophoresed in 0.5X TBE buffer for 2.5 hours at 80V. A molecular weight standard marker XIII (Boehringer Mannheim, size ranging from 2642bp to 50bp) was also loaded for comparative evaluation with the size of the SSR amplicons (i.e., alleles). Gels were stained with an ethidium bromide solution for 15 minutes, and then de-stained in distilled, deionized water for 15 minutes. The gels were exposed under Ultra-Violet light and the banding images were captured by an image analysis system (GelDoc2000, BioRad, Hercules, CA) and printed on thermal-sensitive photography paper. This intensity of the ethidium bromide stain was used as a guide to help equalize the sample DNA concentrations among samples.

Phenotypic Markers

Several vegetative and reproductive pigmentation traits, and the genes controlling these traits, were segregating in these three populations. Such segregating genes, when scored on all progenies of the population, can also be used as markers (Table 1).

Williams 82 has black hila on yellow seed coats, tawny pubescence, tan pods, and white flowers. U06-103459 has buff hila on yellow seed coats, grey pubescence, tan pods, and white flowers, while RMLPC1-311-128-128 has imperfect black hila on yellow seed coats, grey pubescence, brown pods, and purple flowers. Pubescence color is controlled by a single locus *T* on LG-C2 (Chr 6), where *TT* and *Tt* genotypes are tawny and *tt* is grey. Pod color is controlled by two loci, each with dominant alleles, on LGs L (Chr 19) and N (Chr 3), where the two-locus homozygotes of $L_1L_1L_2L_2$ or $L_1L_1l_2l_2$ are black, but $l_1l_1L_2L_2$ is brown, $l_1l_1l_2l_2$ is tan. Because all three parents in this study were l_1l_1 homozygotes, segregation was limited to the brown (L_2L_2 or L_2l_2) and tan (l_2l_2) pod colors. Flower color is controlled by six genes: *W1, W2, W3, W4, Wm,* and *Wp* (Palmer et al., 2004 and Takahashi et al., 2008), but all three parents were identical homozygotes for all of these loci, except for the *W1* locus, where *W1W1* and *W1w1* are purple and *w1w1* is white. These phenotypic markers were also used to confirm the purity of parent lines. For those populations in which the parents differed at epistatic pigmentation marker loci (i.e., *R/r*), the $F_{2.4}$ progeny rows were phenotyped to determine is segregation F_2 genotypes were present for those loci.

SNP Marker Analysis

DNA samples, adjusted for concentration, of the $F_{2.4}$ progenies that were in the highest and lowest decile phenotypic classes (i.e., 45 high and 45 low seed oil) plus two DNA samples of each parent and two F_1 samples (F1A and F1B) were transferred to a 96-well reaction plate. Each DNA sample was 50µl with at least 100ng/µl concentration.

The DNA samples of each parent, and their hybrid F1A, and F1B DNA samples were placed in the middle of the 96-well plate, to reduce the probability of genotyping failures (which are more probable at the plate edges).

In this study, the GoldenGate assay and the Illumina® Genotyping Platform (Illumina Inc., San Diego, CA) were used for the SNP genotyping of $F_{2.4}$ progenies in the low and high decile groups, the two parents, and two F_1 individuals. A GoldenGate assay had been developed for 1536 SNP markers that were distributed over the lengths of the 20 chromosomes of the soybean genome (Hyten et al., 2010) (Fig. 2). The genotyping of the 96 DNA samples of each population for the 1536 SNPs was conducted and completed over a 3-day period by Dr. Perry Cregan's staff at his USDA-ARS laboratory at Beltville, MD in March of 2011. The first day consisted of (i) making activated DNA, (ii) adding DNA to oligonucleotides and hybridize, and (iii) extending, ligating, and cleaning up the product, and finally performing the (iv) universal PCR cycle at 1536-plex. The second day consisted of (i) binding PCR product, eluting the dye-labelled strand, and preparing for hybridization, and then (ii) hybridizing to the Sentrix® Array Matrix or BeadChip. The third day consisted of (i) washing and drying the Array Matrix or BeadChip, and then (ii) imaging Array Matrix or BeadChip (Illumina, 2009). The SNP detection analysis was performed at the Soybean Genomics and Improvement Laboratory, USDA-ARS, BARC-West, Beltsville, MD.

Not all of the 1536 SNP loci were expected to be parentally polymorphic in each population, but about 300 to 500 (20-30%) of 1536 SNPs were expected, based on past experience, to be segregating in each of three populations because of the genomic diversity between each pair of mated parents. Because this is a tri-parent mating set, and

recognizing that SNP loci are bi-allelic, a SNP locus polymorphic in one population is expected to be polymorphic in a second population, but is not expected to be polymorphic in the third population. The automatic allele calling for each locus was accomplished in Dr. James E. Specht's lab by using Genome Studio software (Illumina Inc., San Diego, CA). Genome Studio calls of "A" and "B" are based on the homozygous SNP genotype fluorescence output signals of the colors, red and green, respectively. The "H" call was used for heterozygous fluorescence signal of intermediate color. A dash (-) was assigned if an F_2 individual had no fluorescence signal (i.e., denoting a missing genotype for a given SNP). Ultimately, the fluorescence signal-based A and B genotype coding had to be converted to parental-based A and B genotype coding (i.e., all SNP alleles from one parent are assigned A; those from the other parent are assigned B). In the two populations (i.e., UX2428 and UX2430) involving Williams 82 as one parent, Williams 82 was arbitrarily made the A genotype parent for all SNP loci. In the other population (i.e., UX2427) not involving Williams 82, parent line RMLPC1-311-128-128 was arbitrarily assigned the A genotype parent.

Data Analysis

To conduct the QTL mapping analysis using the R/qtl software (Broman et al., 2003), an Excel file *.csv file was created for each of the three populations. In the *.csv file of a given population, the first Excel row was used for the individual F_2 ID number. The ID number was not necessarily contiguous, because some F_2 individuals were not available (i.e. F2 plant died or else produced insufficient F_3 seed to advance). The second

Excel row was a contiguous set of assigned numbers from the first to the last individual of a given population. This contiguous number assignment provided an index number needed for some analytic components of the R/qtl software. The third to fifth rows were the first replication phenotypic data of $F_{2.3}$ progeny seed protein, oil, and moisture content, respectively, and the sixth to eighth rows were the second replication phenotypic data of $F_{2.3}$ progeny seed protein, oil, and moisture content, respectively. The ninth to eleventh rows were used for the mean values of the forgoing first and second replications of the $F_{2.3}$ phenotypic data. The twelfth to fourteenth rows were the raw $F_{2.4}$ seen protein, oil and moisture phenotypic data output directly from the NIR. The single replicate raw values for the $F_{2.4}$ progenies were adjusted for the field block effects using replicated block parent data, the adjusted $F_{2.4}$ seed protein and oil were the fifteenth and sixteenth Excel rows. The seventeenth Excel row was used to denote the mating direction (i.e., F = forward or R = reciprocal) source of each F_2 individual, and the eighteenth Excel row was used for the F_1 plant number source of each F_2 individual, which could be used to denote the F_1 progenitor of the families of F_2 plants in the 90 select genotyped. The remaining rows in the *.csv file contained a genotype of A, H, or B for each F_2 individual for the given SNP marker. There were a total of 571 SNP markers plus any pigmentation phenotypic markers (i.e., just those segregating in the given population), and these were arranged first by chromosome (1 to 20) and then by their currently known chromosomal position. The chromosomal marker positions (cM) published by Hyten et al. (2010) were also included after the SNP marker names.

Broad-sense heritability (H^2) of seed protein and oil content was computed using phenotypic data collected on parents and F_2-derived F_4 seed progenies. For heritability

calculations the environmental variance (σ_e^2) for seed oil content was calculated from parental oil distributions, using this formula;

$$\sigma_e^2 = \frac{1}{2}(\sigma_{mp}^2 + \sigma_{fp}^2)$$

where σ_{mp}^2 is the phenotypic variance of the selfing progenies obtained from homozygous male parent plants and σ_{fp}^2 is the phenotypic variance of the selfing progenies obtained from homozygous female parent plants. The genotypic variance (σ_g^2) of the $F_{2.4}$ progenies was estimated by subtracting the estimated environmental variance from the $F_{2.4}$ phenotypic variance, using this formula;

$$\sigma_g^2 = \sigma_p^2 - \sigma_e^2$$

where σ_p^2 is the phenotypic variance of the $F_{2.4}$ progenies in the given population, and σ_e^2 is the environmental variance estimated as noted above. Broad sense heritability (H^2) was then estimated using the following formula:

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2}$$

Note that the estimated genetic variance used in the above formula includes the additive, dominance, and epistatic components and that these are estimated using F_2-derived F_4 phenotypic data.

Linkage Mapping and QTL Analysis

A Chi-square test was used to identify SNP markers with segregation distortion. A single Chi-square test significance criterion would be $\alpha = 0.05$, but this would not be a suitable criterion for the present case. A genome-wide test criterion was obtained by dividing 0.05 by the number of markers segregating in the given population.

In this study, QTL mapping was performed using R/qtl software. The marker order and Kosambi map distances of the soybean genetic map Version 4.0 reported by Hyten et al. (2010) were used in the population *.csv file that was input into R/qtl for the QTL analysis. The R/qtl "suspect.markers" command was used to find markers whose Chi-square test value for genotypic segregation ratio differed significantly from the expected 1:2:1 (A:H:B) ratio. The R/qtl "drop.markers" command was used to drop those suspect markers. The R/qtl "errorlod" command was used to detect potential genotyping errors. Based on the errorlod list, markers identified as having potential genotyping errors were re-examined with respect to genotypic A:H:B separation graph output from the Genome Studio software (Illumina Inc., San Diego, CA). Markers whose A-H-B genotypes were not clearly separated into distinct clusters were dropped. The R/qtl "countXO" command was used to identify individuals with an excessive total number of crossovers. Such individuals were likely not authentic members of the given population and were removed. To determine if the input Hyten marker order was a good fit for each population, the R/qtl "ripple" command (method = XO) was used to identify the best marker order for each chromosome. The ripple results for each chromosome were then compared to the Williams 82 chromosomal marker sequence order in SoyBase (2011). The final marker order for each chromosome, when the Hyten et al. (2010) and F2

marker orders differed, was settled by examing the marker position in the genome sequence.

QTLs were first detected using simple marker regression (MR) and then with interval mapping using both the expectation maximization (i.e., EM) and the imputation methods (i.e., IMP). Seed oil phenotypes were available for all $F_{2.4}$ progenies (about 500) per population, but SNP marker genotypes were available only for the progenies in the selectively genotyped decile tails. However, genotypes for the phenotypic markers of flower, pubescence, and pod color were obtained for all progenies. Missing genotypes were coded with a dash.

To ascertain the statistical significance of the LOD score peaks (i.e., putative QTLs), 1900 permutation tests were conducted to generate a population-specific genome-wise LOD score significant criterion for each LOD score scan for putative QTLs. Additive and dominant effects of each QTL were estimated from the AA:AB:BB genotypic values for the SNP marker most closely linked to the QTL of interest, whose alleles due to linkage were expected to also be AA:AB:BB.

RESULTS AND DISCUSSION

F_1 Hybridity Confirmation and Development

The putative F_1 plants were evaluated with SSR markers to ensure they were hybrid and not female parent selfs. Based on the initial screen of several SSR markers, eight markers were found to be parentally polymorphic for one or more of the three populations. Four of those eight markers (Satt126, Satt173, Satt309, and Satt565) were parentally polymorphic for UX2427; one of the eight markers (Satt673) was parentally polymorphic for UX2428; and all eight of the markers (Satt126, Satt173, Satt309, Satt345, Satt424, Satt565, Satt589, and Satt673) were parentally polymorphic for UX2430. Ultimately, three markers (Satt126 for UX2427; Satt673 for UX2428: and Satt673 for UX2430) were chosen to conduct the F_1 plant hybridity confirmation (Table 2). Using these markers, the total number of F_1 plants confirmed as hybrid in populations UX2427, UX2428, and UX2430 were eight, 11, and 12, respectively (Table 2). The total number of F_2 seeds obtained from these F_1 plants in populations UX2427, UX2428, and UX2430 were 524, 485, and 508, respectively.

With respect to population development, two of three populations (UX2428 and UX2430) were developed by the mating of high oil lines RMLPC1-311-128-128 and U06-103459 with cultivar Williams 82, while the other population (UX2427) was created by the mating of the two high oil lines to each other. Because the two high oil parents were not released pure lines, there was some SNP locus heterogeneity detected in these two parents. In hindsight, it would have been better to select a single plant of each of these parents to use as both a pollen donor and pollen recipient.

Phenotypic Data Analyses of Parents

Based on 2010 summer field data, the seed oil content of three parents averaged over blocks ranged from 192.2 to 215.3 g kg^{-1} (Table 3). The low oil parent Williams 82 had a mean oil content of 192.2 g kg^{-1} and a standard deviation of 4.3 g kg^{-1}. High oil parents RMLPC1-311-128-128 and U06-103459 had mean oil contents of 215.3 and 211.2 g kg^{-1}, respectively, and standard deviations of 3.1 and 4.1 g kg^{-1}, respectively.

The seed protein content of three parents harvested along with $F_{2.4}$ progenies ranged from 308.1 to 358.0 g kg^{-1} (Table 3). The low oil parent Williams 82 had the highest protein content of 358.0 g kg^{-1} and a standard deviation of 6.8 g kg^{-1}.The two high oil parents, RMLPC1-311-128-128 and U06-103459, had lower mean protein contents of 308.1 and 332.5 g kg^{-1}, respectively, and had standard deviations of 6.2 and 7.8 g kg^{-1}, respectively.

The standard protein and oil content values of Williams 82 were published in National Genetic Resources Program (NGRP) soybean germplasm database. The standard values of high oil parent U06-103459 were based on 2007 UNL high oil tests, and the standard values of parent line RMLPC1-311-128-128 were based on 2005 Nebraska soybean variety tests. The aforementioned values are presented in Table 1. The phenotypic data of this thesis study was measured with a 13% (130 g/kg seed) moisture basis (Table 3), while the measurements made in 2005 and 2007 were based on a 0% basis. The following is the moisture basis conversion formula:

$$P_2 = \left(\frac{100 - M_2}{100 - M_1}\right) P_1$$

Here, M_1 is the original moisture, M_2 is the new moisture basis, P_1 is the original constituent percentages (under M_1), and P_2 is the adjusted constituent percentages at moisture M_2. Therefore, all the measurements of protein and oil content in 2005 and 2007 were converted into a 13% (130 g/kg seed) of moisture basis for comparison convenience (Table 1).

There were a slight differences between the parental seed protein and oil values measured in this thesis study compared to the standard values. It was observed that the two high oil parents RMLPC1-311-128-128 and U06-103459 had lower seed protein and lower seed oil values (Table 3) compared to the standard values (Table 1). One possible reason is that the measurements of standard values of these two parental lines were based on one-cup NIR seed samples back then, which may result in less precise measurements compared with full-cuvette NIR seed samples used now. In contrast, compared to the standard values reported in NGRP database (Table 1), Williams 82 had slightly higher seed protein and higher oil content values in this study (Table 3).

Phenotypic and Genotypic Data Analyses

Phenotypic Correlations

Soybean seed protein and oil content have been found frequently highly to be negatively correlated (Burton, 1987). In this thesis study, negative phenotypic correlations between seed protein and oil were observed in all three $F_{2.4}$ populations based on statistically significant test of $\alpha = 0.05$. The correlation observed in these three populations ranged from r = -0.78 to -0.70. Population UX2428 had the highest negative

correlation (-0.78), though population UX2427 and UX2430 also had strong negative correlations between seed protein and seed oil of -0.70 and -0.72 respectively. Figure 3 clearly shows the highly negative correlation between seed protein and seed oil in all three populations. Two explanations for the negative phenotypic association between these two traits observed at the genotypic level are typically hypothesized. One is that this inverse association is caused by two tightly linked QTLs; one QTL controlling *only* protein content and the other QTL controlling *only* oil content, but with the high protein allele at the protein QTL and the low oil allele for the oil QTL (or *vice-versa* alleles) locked via tight linkage in a repulsion phase. The alternative hypothesis is just one single QTL with coincident pleiotropic control over both protein and oil, such that there exists an allele for high protein and low oil and the contrasting allele with low protein and high oil. These hypotheses can be evaluated when QTLs are identified – see the subsequent QTL detection section.

Broad Sense Heritability

In most of the situations, heritability estimates are usually not determined without replications in both space (i.e., different locations) and time (i.e., different years). In this study, however, the broad sense heritability estimates were based on single one-field one-year replicates and simply computed to conduct comparisons among the three populations. The broad sense heritability of seed oil content computed for each of these three populations indicated that some of the phenotypic variation was likely genetic. The seed oil heritability of these three populations ranged from 44 to 58%. Population UX2427 had the seed oil heritability of 47%, and UX2428 and UX2430 had seed oil heritability of 58 and 44%, respectively. Although the smallest seed oil content difference

(2.5 g gk^{-1}) between two parents was observed in UX2427, it was of interest to note that the heritability of seed oil content of population UX2427 was comparatively higher than the seed oil content heritability of UX2430.

Seed protein heritability of the three populations ranged from 50 to 68%. The heritability of populations UX2427, UX2428, and UX2430 were 50, 68, and 54%, respectively. According to several published papers, seed protein heritability is quite high (Brummer et al., 1997; Chung et al., 2003). Nevertheless, Brim and Burton (1979) reported that the seed protein heritability could be as low as 20 to 39%. However, in their study, the seed protein heritability estimates were estimated based on the matings of parents that had only slight difference in seed protein content.

Mating Direction

Gilsinger et al. (2010) reported that maternal effects were apparently important in the inheritance of the fatty acid composition of soybean seed. In the present study, there were two mating directions (i.e., forward and reciprocal) for each population (Table 2). However, there was no statistically significant difference between two mating directions based on the phenotypic data analysis conducted in the present study. Figure 4 shows the box plots of seed oil content of two mating directions. Gilsinger et al. (2010) suggested that when using two parents with only little differences in oil content, the maternal effects could be easily masked by environmental variance. It might be better if the two parents had larger differences in seed oil content, as that could increase the ability to detect significant maternal effects.

Phenotypic and Genotypic Data Analyses of Progenies

The phenotypic and genotypic data of the three populations of this thesis study will be discussed separately. Population UX2430 and UX2428 will be discussed first because they were developed as crosses of different high oil lines with a low oil line (i.e., Williams 82), and population UX2427 will be discussed at the end because it was created as a cross of the two high oil lines.

UX2430

Progeny Data

The seed oil values of the 373 $F_{2.4}$ progenies of population UX2430 exhibited continuous variation (Fig. 5a). Although the distribution was slightly leftward skewed (-0.22) and somewhat leptokurtic – a word meaning a more acute peak and "fatter" tails (0.34), the progeny seed oil distribution was still found to be normally distributed based on the non-significant value (Pr > 0.05) computed for the Shapiro-Wilk (SW) test (Table 4). The parental means indicated that the low oil parent Williams 82 (192.2 g kg^{-1}) and the high oil parent U06-103459 (211.2 g kg^{-1}) differed by 19 g kg^{-1} in seed oil content (Table 3). The seed oil content mean for the progeny was 200.0 g kg^{-1}, which was very close to the mid-parent mean of 201.7 g kg^{-1}. The 373 progeny oil values ranged from 181.4 to 211.4 g kg^{-1}(Table 4), and did not differ too much from the parental ranges (Table 3).

The focus of this thesis research was the identification of seed oil QTLs whose alleles had little or no inverse impact on seed protein. Therefore, the progeny seed protein values are also presented here. The seed protein content of the UX2430 $F_{2.4}$ population also showed continuous variation. The normality of the distribution was confirmed by the

non-significant SW test value (Table 5). The low oil but high protein parent Williams 82 (358.0 g kg^{-1}) and the high oil but low protein parent U06-103459 (332.5 g kg^{-1}) differed by 25.5 g kg^{-1} in seed protein content (Table 3). The seed protein content mean for the progeny ranged was 347.0 g kg^{-1}, which was also close to the mid-parent mean of 345.3 g kg^{-1}. The 373 progeny protein values (Table 5) ranged from 316.5 to 387.3 g kg^{-1}, and did not differ too much from the parental ranges (Table 3).

Classical Marker Genotypes

The pigmentation phenotypes of low oil parent Williams 82 are white flowers, tawny pubescence and black hila, so its genotype is *w1w1TTRR*. The pigmentation phenotypes of the high oil parent U06-103459 are white flowers, grey pubescence, and buff hila, so its genotype is *w1w1ttRR* or *w1w1ttrr*. If U06-103459 is *w1w1ttRR*, then just black and buff hila colors would be observed from the progenies; if it is *w1w1ttrr*, black, buff, and brown hila would be observed. Inspection of the progeny hila colors revealed no brown hila colors, so the pigmentation genotype of U06-103459 was confirmed as *w1w1ttRR*.

Linkage Map Analyses

In population UX2430, 570 of the 1536 SNP plus the *T* locus (Chr 6) for the pubescence color were presented in the *.csv file that was imported to R/qtl in order to construct a genetic linkage map based on the marker order of the soybean integrated genetic linkage map (Consensus Map 4.0) published by Hyten et al. (2010). Hyten's genetic linkage map spans 2296.4 cM of Kosambi map distance (Fig. 2). The map used for that locus was that shown in SoyBase (2011), which is at 101.5 cM on Chr6. The

R/qtl "suspect.makers" command was used to identify SNP markers with significant segregation distortion, which was judged by $P < 0.0001$ (i.e., $1e^{-4}$). The R/qtl "drop.markers" command was used to drop these 224 SNP markers, plus three other SNP markers that had either a high errorlod score and/or a poor separation of A/H/B genotypes in GenomeStudio output. The 343 SNP markers that remained were used to construct a new genetic linkage map of UX2430 $F_{2.4}$ population. Unfortunately, 110 F_1-derived F_2 individuals of family #10 (i.e., the second reciprocal family) had to be dropped because this F_1-derived family exhibited a recombination of pattern that differed from the nine other families. Ultimately, 343 markers and 373 F_2 individuals remained for constructing the UX2430 F_2 plant genetic linkage map. For a final genetic map, F_2-specific marker-to-marker recombination fraction values were computed. This resulted in a large gap in the Chr 1 map. Because of an absence of markers in this region, the recombination fraction was greater than 0.49 (= 115 cM in Kosambi). The relationship between recombination fraction and map distance is exponential at recombination fraction values exceeding 0.490. The R/qtl "fix" command was used to bring large map distances to a more suitable 115 cM gap subsequent QTL LOD score scanning purposes. Finally, the R/qtl "ripple" command was run to finalize the marker order in each of the 20 chromosomes. Fig. 6a displays the final UX2430 genetic map. There is some expansion in the F2 map compared with Hyten linkage map. This is due to fewer markers (larger gaps) and fewer individuals (Fig. 7a).

QTL Mapping Analysis

In this thesis study, three LOD score scans were conducted to detect significant protein and oil QTLs: single marker regression analysis (i.e., MR method) and two

interval analyses: one using the Expectation Maximization (i.e., EM) Algorithm (i.e, maximum likelihood), and the Multiple Imputation (i.e., IM) method, which allows simple ANOVA to be performed. The EM method has been frequently used when analyzing selective genotyping data. The IM method has been used when dealing with *random* missing genotype data, but in selective genotyping, marker genotypes are *purposefully* missing. Although all three methods were used in the analyses, only the EM method was suitable for selective genotyping QTL detection, so only the results of that method are reported here.

Two oil QTLs were detected with the EM method in UX2430 population that had a LOD scores ≥ 3.0 (Table 6), and both of these two QTLs were statistically significant based on the 95th percentile of genome-wide maximum LOD score (i.e., 3.66 in this population) that was generated with 1900 permutations (Fig. 8a). The SNP markers nearest to these two QTLs were S12382 and S10452, which were located at 138.2 cM on LG-C2 (Chr 6) and 66.5 cM on LG-M (Chr 7), respectively. Table 7 shows the flanking markers of each statistically significant QTL based on the Bayes Credible Interval computation (i.e., a type of confidence interval, C.I.). These two QTLs explained 7.0% and 4.9% of variation for seed oil, respectively (Table 6), which in effect are independent heritability estimates specific for the QTLs. The high oil parental line U06-103459 S12382 allele had positive additive effect, indicating that this parental allele was associated with high seed oil, whereas U06-103459 S10452 allele had negative additive effect (Fig. 9a).

Three protein QTLs were detected with EM method that had a LOD score ≥ 3.0 (Appendix Table 1). However, only two were confirmed to be statistically significant

based on permutation determined genome-wide LOD score (i.e., 3.57) (Appendix Fig. 1a). The nearest markers were S16994 and S10452, which were located at 122.2 cM on LG-C2 (Chr 6) and 65.3 cM on LG-M (Chr 7), respectively. The high oil but low protein parental line U06-103459 S16994 allele was found to have negative additive effect on seed protein (-4.0 g kg^{-1}), whereas U06-103459 S10452 allele was found to have positive additive effect on seed protein (3.2 g kg^{-1}) (Appendix Fig. 2a), and each explained 8.3% and 5.2% of variation, respectively.

In population UX2430, one SNP marker S10452 was associated with a QTL governing with seed oil and seed protein, so the high oil parent U06-103459 S10452 allele was associated with both high oil and low protein, whereas the allele from low oil parent Williams 82 was associated with both low oil and high protein. This allelic phenotypic scenario is consistent with the well-known genotypic level negative correlation between soybean seed protein and seed oil content.

UX2428

Progeny Data

The seed oil values of the 389 $F_{2.4}$ progenies of population UX2428 exhibited continuous variation (Fig. 5b). Although the distribution was slightly rightward skewed 0.12) and platykurtic (-0.08), the progeny seed oil distribution was still found to be normally distributed based on the non-significant value (Pr > 0.05) computed for the Shapiro-Wilk (SW) test (Table 4). The parental means indicated that the low oil parent Williams 82 (192.2 g kg^{-1}) and the high oil parent RMLPC1-311-128-128 (215.3 g kg^{-1}) differed by 23.1 g kg^{-1} in seed oil content (Table 3). The seed oil content mean for the

progeny was 205.1 g kg^{-1}, which was very close to the mid-parent mean of 203.8 g kg^{-1}. The 389 progeny oil values ranged from 189.6 to 224.1 g kg^{-1}(Table 4), and did not differ too much from the parental ranges (Table 3).

The seed protein content of the UX2428 $F_{2.4}$ population also showed continuous variation. The normality of the distribution was confirmed by the non-significant SW test value (Table 5). The low oil but high protein parent Williams 82 (358.0 g kg^{-1}) and the high oil but low protein parent RMLPC1-311-128-128 (308.1 g kg^{-1}) differed by 49.9 g kg^{-1} in seed protein content (Table 3). The seed protein content mean for the progeny ranged was 330.0 g kg^{-1}, which was also close to the mid-parent mean of 333.1 g kg^{-1}. The 389 progeny protein values (Table 5) ranged from 295.4 to 366.4 g kg^{-1}, and did not differ too much from the parental ranges (Table 3).

Classical Marker Genotypes

The pigmentation phenotypes of low oil parent Williams 82 are white flowers, tawny pubescence and black hila, so its genotype is *w1w1TTRR.* The pigmentation phenotypes of the high oil parent RMLPC1-311-128-128 are purple flowers, grey pubescence, and imperfect black hila, so its genotype is *W1W1ttRR.* Therefore, it is clear that the progenies would have been segregating in both *W1* and *T* loci, which means there would be black (*W1TR* and *w1TR*), imperfect black (*W1tR*), and buff (*w1tR*) hila observed in the progenies.

Linkage Map Analyses

In population UX2428, 570 of the 1536 SNP plus the *T* locus (Chr 6) for the pubescence color, L_2 locus (Chr 3) for the pod color, and *W1* locus (Chr 13) for the

flower color were presented in the *.csv file that was imported to R/qtl in order to construct a genetic linkage map based on the marker order of the soybean integrated genetic linkage map (Consensus Map 4.0) published by Hyten et al. (2010) (Fig. 2). The map used for those three pigmentation loci were that shown in SoyBase (2011), which is at 101.5 cM on LG-C2 (Chr 6), 2.7 cM on LG-N (Chr 3) and 19.2 cM on LG-F (Chr 13), respectively. The R/qtl "suspect.makers" command was used to identify SNP markers with significant segregation distortion, which was judged by $P < 0.0001$ (i.e., $1e^{-4}$). The R/qtl "drop.markers" command was used to drop these 248 SNP markers, plus four other SNP markers that had either a high errorlod score and/or a poor separation of A/H/B genotypes in GenomeStudio output. The 321 SNP markers that remained were used to construct a new genetic linkage map of UX2428 $F_{2.4}$ population. 41 F_1-derived F_2 individuals of family #1 and 38 F_1-derived F_2 individuals of family #2 had to be dropped because exhibited a recombination of pattern that differed from the eight other families. In addition, five F_2 individuals were dropped because of excessive crossover number. Ultimately, 321 markers and 389 F_2 individuals were used to construct the $F_{2.4}$ genetic linkage map. For a final genetic map, F_2-specific marker-to-marker recombination fraction values were computed. This resulted in large gaps in the map of Chr 7, 11, and 13. The R/qtl "fix" command was used to bring large map distances to a more suitable 115 cM gap for subsequent QTL LOD score scanning purposes. Finally, the R/qtl "ripple" command was run to finalize the marker order in each of the 20 chromosomes. Fig. 6b displays the final UX2428 genetic map. There is some expansion in the F_2 map compared with Hyten linkage map because of fewer markers (larger gaps) and fewer individuals (Fig. 7b).

QTL Mapping Analysis

Three oil QTLs were detected with the EM method in UX2428 that had LOD scores ≥ 3.0 (Table 6), and all three were statistically significant based on the 95th percentile of genome-wide maximum LOD score (i.e., 3.62) that was generated with 1900 permutations (Fig. 8b). The SNP markers nearest to these three oil QTLs were S17276, S12243, and S01447, and the map positions of these three oil QTLs were located at 287.1 cM on LG-F (Chr 13), 117.7 cM on LG-E (Chr 15), and 122.2 cM on LG-L (Chr 19), respectively. These QTLs explained 4.4%, 4.8%, and 7.7% of the variation, respectively (Table 6). All of RMLPC1-311-128-128 alleles had positive additive effects on seed oil content (Fig. 9b). Table 9 shows the flanking markers of each statistically significant seed oil QTL based on the Bayes C.I. analysis.

Four seed protein QTLs were detected with EM method that had LOD scores ≥ 3.0 (Appendix Table 1). These QTLs were located on LGs D1b (Chr 2), C2 (Chr 6), B1 (Chr 11), and L (Chr 19). However, only two of the four protein QTLs were statistically significant based on permutation determined genome-wide LOD scores (i.e., 3.62) (Appendix Fig. 1b). The SNP markers nearest to these two protein QTLs were located at 158.0 cM on LG-C2 (Chr 6) and 132.9 cM on LG-L (Chr 19), and they explained 4.9% and 5.1% of variation in seed protein, respectively. The negative additive effects were detected with both high oil low protein parent RMLPC1-311-128-128 alleles (Appendix Fig. 2b).

Comparing the positions of detected protein and oil QTLs, there were two seed oil QTLs, which the nearest SNP markers were S17276 on LG-F (Chr 13) and S12243 on

LG-E (Chr 15), had no corresponding seed protein QTLs. This indicated that these two seed oil QTLs may have only slight or no impact on seed protein content.

UX2427

Progeny Data

The seed oil values of the 490 $F_{2.4}$ progenies of population UX2427 exhibited continuous variation (Fig. 5c). Although the distribution was slightly rightward skewed 0.13) and platykurtic (-0.02), the progeny seed oil distribution was still found to be normally distributed based on the non-significant value (Pr > 0.05) computed for the Shapiro-Wilk (SW) test (Table 4). The parental means indicated that the high oil parent RMLPC1-311-128-128 (215.3 g kg^{-1}) and the other high oil parent U06-103459 (211.2 g kg^{-1}) differed by only 4.1 g kg^{-1} in seed oil content (Table 3). The seed oil content mean for the progeny was 213.5g kg^{-1}, which was very close to the mid-parent mean of 213.3 g kg^{-1}. The 490 progeny oil values ranged from 199.7 to 229.1 g kg^{-1}(Table 4), and did not differ too much from the parental ranges (Table 3).

The seed protein content of the UX2427 $F_{2.4}$ population also showed continuous variation. The normality of the distribution was confirmed by the non-significant SW test value (Table 5). The high oil low protein parent RMLPC1-311-128-128 (308.1 g kg^{-1}) and the other parent U06-103459 (332.5 g kg^{-1}) differed by 24.4 g kg^{-1} in seed protein content (Table 3). The seed protein content mean for the progeny ranged was 324.0 g kg^{-1}, which was also close to the mid-parent mean of 320.3 g kg^{-1}. The 490 progeny protein values (Table 5) ranged from 291.6 to 353.3 g kg^{-1}, and did not differ too much from the parental ranges (Table 3).

Classical Marker Genotyping

The pigmentation phenotypes of the high oil parent RMLPC1-311-128-128 are purple flowers, grey pubescence, and imperfect black hila, so its genotype is *W1W1ttRR.* The pigmentation phenotypes of the other high oil parent U06-103459 are white flowers, grey pubescence, and buff hila, so its genotype is *w1w1ttRR*, which was confirmed in population UX2430. It is needed to note that it was not able to determine the U06-103459 genotype in this population. If U06-103459 is *w1w1ttRR*, imperfect black and buff hila would be observed; if it is *w1w1ttrr*, there would still only black and buff hila observed. Therefore, it is clear that the progenies would have been segregating in *W1* locus, which means there would be imperfect black (*W1tR*) and buff (*w1tR*) hila observed in the progenies.

Linkage Map Analyses

In population UX2427, 572 of the 1536 SNP plus the L_2 locus (Chr 3) for the pod color and *W1* locus (Chr 13) for the flower color were presented in the *.csv file that was imported to R/qtl in order to construct a genetic linkage map based on the marker order of the soybean integrated genetic linkage map (Consensus Map 4.0) published by Hyten et al. (2010) (Fig. 2). The map used for these two locus was that shown in SoyBase (2011), which are located at 2.7 cM on LG-N (Chr 3) and 19.2 cM on LG-F (Chr 13), respectively. The R/qtl "suspect.makers" command was used to identify SNP markers with significant segregation distortion, which was judged by $P < 0.0001$ (i.e., $1e^{-4}$). The R/qtl "drop.markers" command was used to drop these 199 SNP markers, plus 11 other SNP markers that had either a high errorlod score and/or a poor separation of A/H/B

genotypes in GenomeStudio output. The 364 SNP markers that remained were used to construct a new genetic linkage map of UX2427 $F_{2.4}$ population. 17 F_2 individuals had to be dropped because of excessive crossover number. Ultimately, 364 markers and 490 F_2 individuals remained for constructing the UX2427 F_2 plant genetic linkage map. For a final genetic map, F_2-specific marker-to-marker recombination fraction values were computed. This resulted in large gaps in the map of Chr 5, 6, 10, 11, and 13. The R/qtl "fix" command was used to bring large map distances to a more suitable 115 cM gap for subsequent QTL LOD score scanning purposes. Finally, the R/qtl "ripple" command was run to finalize the marker order in each of the 20 chromosomes. Fig. 6c displays the final UX2427 genetic map. There is some expansion in the F_2 map compared with Hyten linkage map. This is due to fewer markers (larger gaps) and fewer individuals (Fig. 7c).

QTL Mapping Analysis

Five oil QTLs were detected with the EM method in UX2427 population that had a LOD score ≥ 3.0 (Table 6), but only three of them were statistically significant based on the 95^{th} percentile of genome-wide maximum LOD score (i.e., 3.83 in this population) that was generated with 1900 permutation (Fig. 8c). The SNP markers nearest to these three oil QTLs were S06956, S10061, and S02534, and the map positions of these QTLs were located at 62.1 cM on LG-B1 (Chr 11), 153.3 cM on LG-E (Chr 15), and 111.4 cM on LG-L (Chr 19), respectively. Table 7 shows the flanking markers of each statistically significant seed oil QTL based on the Bayes C.I. analysis. These three oil QTLs explained 4.2%, 7.2%, and 5.2% of the variation for seed oil, respectively (Table 6). Two of the three RMLPC1-311-128-128 alleles had positive additive effects, which were

linked to markers S10061 and S02534, while the other RMLPC1-311-128-128 allele, which was linked to marker S06956, was detected with negative additive effect (Fig. 9c).

Four protein QTLs were detected with EM method that had LOD scores ≥ 3.0 (Appendix Table 1). However, only three of them proved to be statistically significant based on permutation determined genome-wide LOD scores (i.e., 3.71), and they were located on LGs B1 (Chr 11), E (Chr 15), and L (Chr 19) (Appendix Fig. 1c). Two of three alleles at markers S10061 and S02534 from RMLPC1-311-128-128 were found to have negative additive effect (-3.0 and -2.1 g kg^{-1}, respectively) (Appendix Fig. 2c), which explained 6.6% and 4.6% of variation respectively, whereas the RMLPC1-311-128-128 S06956 allele were found to have positive additive effect (2.4 g kg-1), which explained 4.5% of variation.

Comparing the positions of detected protein and oil QTLs, all three seed oil QTLs, which the nearest SNP markers were S06956 on LG-B1 (Chr 11), S10061 on LG-E (Chr 15), and S02534 on LG-L (Chr 19), had corresponding seed protein QTLs. This scenario is consistent with the negative correlation between soybean seed protein and seed oil content.

Comparison QTLs Detected with QTLs Previously Reported

Three of the eight statistically significant seed oil QTLs identified by the interval mapping method (EM) were closely located to QTLs that have previously reported by other researchers.

On LG-B1 (Chr 11), SNP marker S06956, which has a Soybean Consensus 4.0 map position of 32.1 cM, is closely linked to a seed oil QTL in the present study. Qi et al. (2011) found that SSR marker Satt251 was associated with seed oil content QTL. The SSR marker Satt251 is located at 38.8 cM on the Consensus 4.0 map. Because there is only 6.7 cM distance between the SNP marker used here and SSR marker used by Qi et al. (2011), it is quite likely that we and they identified the same LG-B1 seed oil QTL.

SNP marker S02534 on LG-L (Chr 19), which was associated with a seed oil QTL in this study, is located at the Consensus 4.0 map position of 88.8 cM. Hyten et al. (2004) found that SSR marker Satt373 on LG-L at position 94.0 cM was also closely linked to a seed oil QTL. The distance between these two markers is 5.2 cM, again suggesting that we and they detected the same oil QTL. SNP marker S01447, which resides at position 90.4 cM on LG-L (Chr 19) is only 3.6 cM from SSR marker Satt373. SNP marker S02534 segregated in UX2427, whereas S01447 segregated in UX2428, but both populations were segregated for the same seed oil QTL.

Theoretically, the oil QTLs detected in this study provide useful information for breeders who will want to select parents for mating that have the desired high oil alleles. Breeders can develop a very high oil breeding line by combining all the possible high oil alleles into one line, and then use the high oil breeding line as a donor parent for developing cultivars suitable for use in biofuel production projects in the future.

Power of QTL Mapping Based on Selective Genotyping

Selective genotyping is an efficient method developed for QTL detection. Instead of genotyping on entire population, only a portion of the population will be genotyped without loss of QTL detection power. Darvasi and Soller (1992) suggested that to maximize the efficiency of selective genotyping, the individuals selected to be genotyped should no more than 50% of entire population (i.e., the highest 25% and the lowest 25%). Ayoub and Mather (2002) reported that, in the North American Barley QTL mapping population, selecting for only 10% or 20% extremes of the total population for genotyping, instead of genotyping the entire population was just as effective in detecting the same QTLs, so decile or quintile selective genotyping is clearly an effective alternative method for QTL detection. In this thesis study, somewhat less than 20% of total population was selected for genotyping in all three populations. There were 17.8% (i.e., 8.9% for each tail), 19.0% (i.e., 9.5% for each tail), and 18.6% (i.e., 9.3% for each tail) of total population selected for genotyping in UX2427, UX2428, and UX2430, respectively. It has been reported that increasing the population size is an effective way to improve the power in selective genotyping (Darvasi, 1993; Kearsey and Pooni, 1996). In the present study, there were approximately 500 individuals in each population, almost double the population sizes that Ritche (2003) used, so the power of QTL detection in this study was relatively high.

CONCLUSIONS

By studying three $F_{2.4}$ mapping populations, some new seed oil QTLs were detected with little or no impact on seed protein content. In total, six different statistically significant seed oil QTLs were identified, and these were located on LGs C2 (Chr 6), M (Chr 7), B1 (Chr 11), F (Chr 13), E (Chr 15), and L (Chr 19). In population UX2428, there were two statistically significant seed oil QTLs, i.e., those on LG-F (Chr 13) and LG-E (Chr 15), for which no significant seed protein QTL was detected at the same or close by map position. However, there have been many protein QTLs detected on LG-E (Chr 15) previously reported. Therefore, only the seed oil QTL located near SNP markers S17276 (Chr 13) is informative, and the high oil parental line RMLPC1-311-128-128 allele at this oil QTL had positive additive effect of 2.1 g kg-1. This allele would likely be of significant interest to soybean breeders working to develop high yielding high seed oil cultivars for producers supplying soybean seed to bio-diesel plants. In contrast, other than this oil QTL, the linkage group regions for which we detected three oil QTLs in UX2427, two oil QTL in UX2428, and two oil QTLs in UX2430 apparently have significant linkage with nearby seed protein QTLs, or are simply not oil QTLs *per se* but are QTLs whose alleles give rise to inverse pleiotropic effects on seed protein and oil content. One of the reasons that population UX2427 was developed, which was the mating of two high oil parents, was because it was expected that if oil QTLs were detected in UX2428 and UX2430, which were the matings of Williams 82 and either of the two high oil parents, those QTLs would not be detected in UX2427 given that QTL allele present in the two high oil parents is assumed to be identical with each other, but different from the allele present in the Williams 82 parent. However, two oil QTLs, one

on LG-C2 (Chr 6) and the other on LG-M (Chr 7) detected only in the UX2430. This is not logically possible, since a QTL detected in one population of a Triallelic mating scheme, ***must*** be detected in at least one of the other two populations. This was also true for population UX2427 in which the seed oil QTL dectected on LG-B1 (Chr 11) was not detected in two other populations. Although each of two populations segregating for the QTL and the third population not segregating is logical qualitative expectation, it is possible that a QTL in one population breached the significance criterion to be declared as such, whereas the same QTL in the other population fall short of statistical significance and was thus ignored. As a result, the oil QTL detected on LG-M (Chr 7) had the high oil parental line RMLPC1-311-128-128 allele with negative additive effect, which means the high oil allele was from the low oil parent Williams 82. This is reasonable that why it could not be detected in UX2427 as Williams 82 was not the parent of this population. One possible reason that the oil QTL detected on LG-C2 (Chr 6) was not found in UX2427 is because the QTL was actually the pubescence color. The *T* gene and *R* gene give a black hilum color, and black hila would affect the reading of NIR analysis. Therefore, it is likely that the QTL detected on LG-C2 (Chr 6) was not an oil QTL.

In conclusion, the results of this thesis research indicated that the seed oil QTL detected on LG-F (Chr13) may be useful for soybean breeders interested in developing high oil breeding lines without lower the seed protein content. If it is true, the SNP S17276 allele from high oil parent RMLPC1-311-128-128 would be worthy to introgress (and thus also drag the QTL high oil alleles linked to those SNP marker alleles) into current high-yield cultivars for use in future industrial soy biofuel production.

Table 1. Parental germplasm descriptions.

Parental name	Maturity group	Seed†		Flower color	Pod color	Pubesence color	Seed coat color	Hilum color
		Protein	Oil					
		-------g kg^{-1}-------						
U06-103459	II	388(338)	250(218)	White	Tan	Grey	Yellow	Buff
RMLPC1-311-128-128	III	361(314)	248(216)	Purple	Brown	Grey	Yellow	Imperfect Black
Williams 82	III	395(344)	208(181)	White	Tan	Tawny	Yellow	Black

† Seed protein and oil values of Williams 82 were published in NGRP soybean germplasm database; the standard values of U06-103459 were based on 2007 UNL high oil tests; and the standard values of RMLPC1-311-128-128 were based on 2005 Nebraska soybean variety tests. The seed protein and oil content values were based on 0% moisture basis while the values in parentheses were based on 13% moisture basis.

Table 2. List of the packet numbers, numbers of seed per packet, seed germination numbers, markers used for F_1 hybridity confirmation, and the hybridity test results obtained in each population.

Population	Parents		Packet Number	Number of seed *per* packet	Number of germination	Marker for F1 confirmation	Imaging score†
	Female	Male		----------No.----------			
UX2427	U06-103459	RMLPC1-311-128-128	1	1	1	Satt126	H
			2	1	1	Satt126	S
			3	1	1	Satt126	H
			4	3	3	Satt126	H
			5	2	2	Satt126	H
UX2427-B	RMLPC1-311-128-128	U06-103459	6	2	1	Satt126	H
UX2428	Williams 82	RMLPC1-311-128-128	1	2	2	Satt673	H
			2	1	0	-	-
			3	3	2	Satt673	H
			4	1	1	Satt673	H
UX2428-B	RMLPC1-311-128-128	Williams 82	5	3	3	Satt673	H
			6	3	3	Satt673	H
			7	3	2	Satt673	S

† Abbreviations: Hybrid and Self.

Table 2. (cont.)

Population	Parents		Packet Number	Number of seed *per* packet	Number of germination	Marker for F1 confirmation	Imaging score
	Female	Male		----------No.----------			
UX2430	U06-103459	Williams 82	1	3	3	Satt673	H
			2	1	1	Satt673	H
			3	3	2	Satt673	H
			4	2	2	Satt673	H
			5	1	1	Satt673	H
			6	2	1	Satt673	H
UX2430-B	Williams 82	U06-103459	7	1	1	Satt673	H
			8	1	1	Satt673	H

† Abbreviations: Hybrid and Self.

Table 3. Seed oil and protein means and other statistical parameters of parental lines obtained in $F_{2.4}$ populations.

Parent	Oil			Protein		
	Mean	Std. Dev.	Max-min	Mean	Std. Dev.	Max-min
	----------g kg-1----------					
U06-103459	211.2	4.1	219-198	332.5	7.8	347-307
RMLPC1-311-128-128	215.3	3.1	222-209	308.1	6.2	328-296
Williams82	192.2	4.3	202-183	358.0	6.8	372-344

Table 4. Seed oil means and other statistical parameters relative to the seed oil phenotypic distributions in the three populations of $F_{2.4}$ progenies.

Population	$F_{2.4}$ Progeny						
	Mean	Std. Dev.	Max-min	Shapiro-Wilk	Pr	Kurtosis	Skewness
	----------g kg-1----------						
UX2427	213.5	4.5	229.1-199.7	1.00	0.89	-0.02	0.13
UX2428	205.1	5.6	224.1-189.6	1.00	0.69	-0.08	0.12
UX2430	200.0	4.8	211.4-181.4	0.99	0.13	0.34	-0.22

Table 5. Seed protein means and other statistical parameters relative to the seed protein phenotypic distributions in the three populations of $F_{2.4}$ progenies.

Population	$F_{2.4}$ Progeny						
	Mean	Std. Dev.	Max-min	Shapiro-Wilk	Pr	Kurtosis	Skewness
	----------g kg-1----------						
UX2427	324.0	9.1	353.3-291.6	1.00	0.54	0.32	-0.15
UX2428	330.0	10.8	366.4-295.4	1.00	0.32	0.25	-0.07
UX2430	347.3	9.9	387.3-316.5	0.99	0.14	0.63	0.14

Table 6. Summary of seed oil QTL peak scores ≥ 3.0, ordered by population, then by chromosome, that were identified by interval mapping using expectation maximization (EM). A permutation test of 1900 replications was conducted in each population to provide a genome-wide 95th percentile LOD score to serve as a statistical significance criterion for evaluating a QTL LOD score peak. The additive (a) and dominant (d) effects were calculated on the basis of the substitution of a high oil parent allele for a low oil parent allele at the given SNP locus. Map position and LOD score values are provided for the corresponding protein QTL for each oil QTL (if applicable).

Pop.		Chr.	LG		LOD	Permutation-based		QTL Effect		Protein QTL	
No.	SNP	No.	name	Position	(if ≥ 3.0)	LOD Score	R^2	a‡	d‡	Pos.	LOD
				cM			%	---g kg-1---		cM	
UX2427	S14594	3	N	42.7	3.23	-	-	-	-	-	-
UX2427	S06956†	11	B1	62.1	4.58	3.83	4.2	-1.2	-0.2	42.1	4.87
UX2427	S02236	13	F	253.7	3.37	-	-	-	-	-	-
UX2427	S10061	15	E	153.3	7.96	3.83	7.2	1.6	-0.7	153.3	7.31
UX2427	S02534†	19	L	111.4	5.72	3.83	5.2	1.4	-0.8	109.4	5.02
UX2428	S17276†	13	F	287.1	3.84	3.62	4.4	2.1	-0.7	-	-
UX2428	S12243	15	E	117.7	4.18	3.62	4.8	1.7	-0.6	-	-
UX2428	S01447	19	L	122.2	6.73	3.62	7.7	1.9	-0.9	132.9	4.40
UX2430	S12382†	6	C2	138.2	5.88	3.66	7.0	1.4	1.4	122.2	7.06
UX2430	S10452	7	M	66.5	4.05	3.66	4.9	-1.6	0.1	65.3	4.32

† nearest marker.

‡ If the effect is negative, the high oil parent marker allele depresses seed oil content.

Table 7. Relative to the data presented in Table 8, shown here are the markers nearest to the left and right 95% Bayes confidence interval (C.I.) with their map positions and oil QTL LOD scores.

			Left boundary of the C.I.			Marker			Right boumdary of the C.I.		
Pop. No.	Chr. No.	LG name	nearest marker	map position	LOD	or Nearest Marker	Map position	LOD‡	nearest marker	map position	LOD
				cM			cM			cM	
UX2427	13	B1	S06956	32.1	3.39	S06956†	62.1	4.58	S07854	147.2	0.02
UX2427	15	E	S02916	140.2	3.97	S10061	153.3	7.96	S06795	185.0	2.64
UX2427	19	L	S05243	86.2	2.80	S02534†	111.4	5.72	S07624	122.9	4.00
UX2428	13	F	S06521	262.2	1.63	S17276†	287.1	3.84	S02236	311.2	2.42
UX2428	15	E	S07592	78.5	1.24	S12243	117.7	4.18	S05112	173.7	0.01
UX2428	19	L	S06809	111.9	4.38	S01447	122.2	6.73	S11193	140.2	4.04
UX2430	6	C2	S16994	122.0	4.87	S12382†	138.2	5.88	S08406	150.2	4.14
UX2430	7	M	S05477	61.3	3.11	S10452	66.5	4.05	S07684	88.1	1.69

† nearest marker.

‡ Only those QTL peaks detected to be statistically significant (determined by using the 95th percentile of genome-wide maximum LOD scores of 1900 permutations) are presented here.

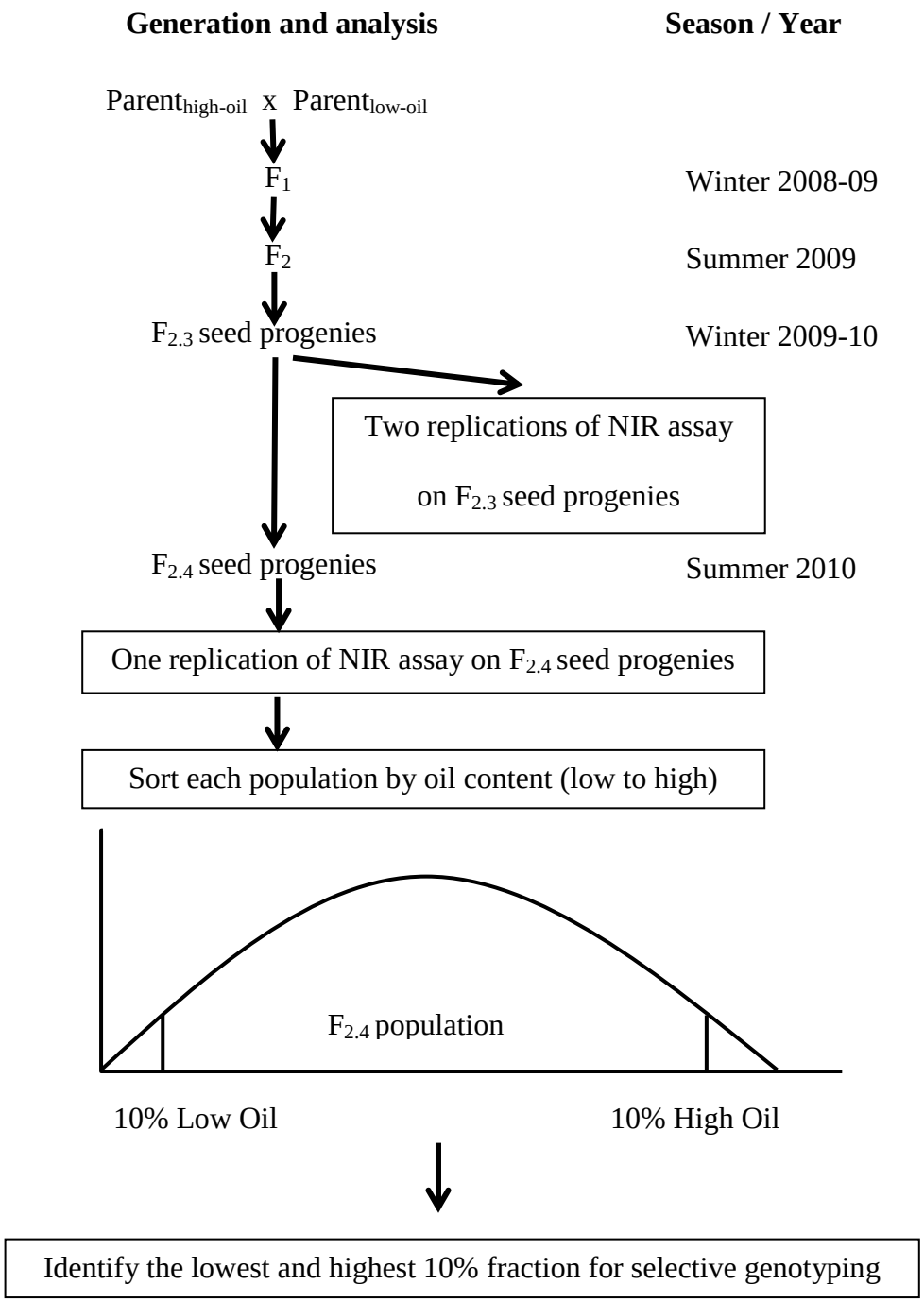


Fig. 1. Development of three F_2 populations and the use of the extreme decile tails of the $F_{2.4}$ seed progeny oil distributions for selective genotyping with 1536 SNP markers.

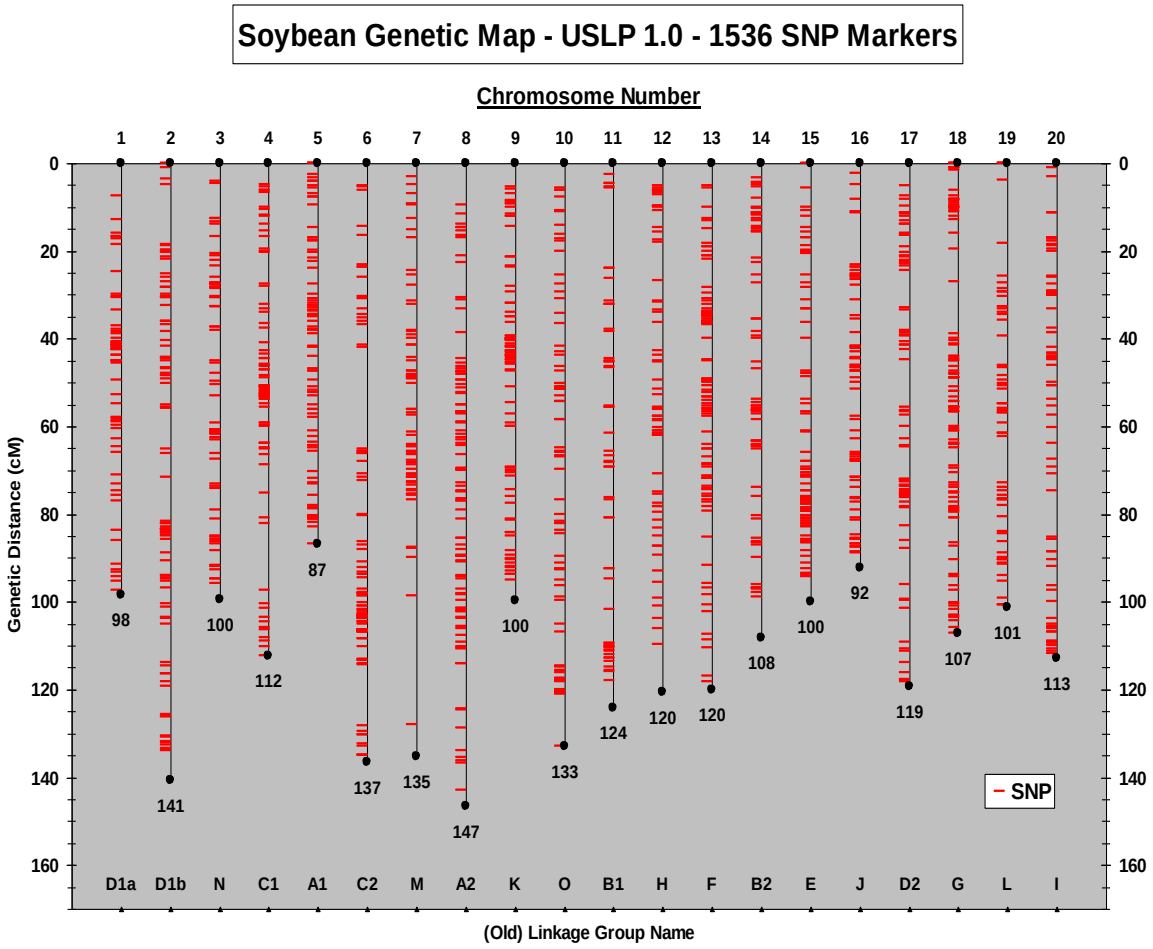
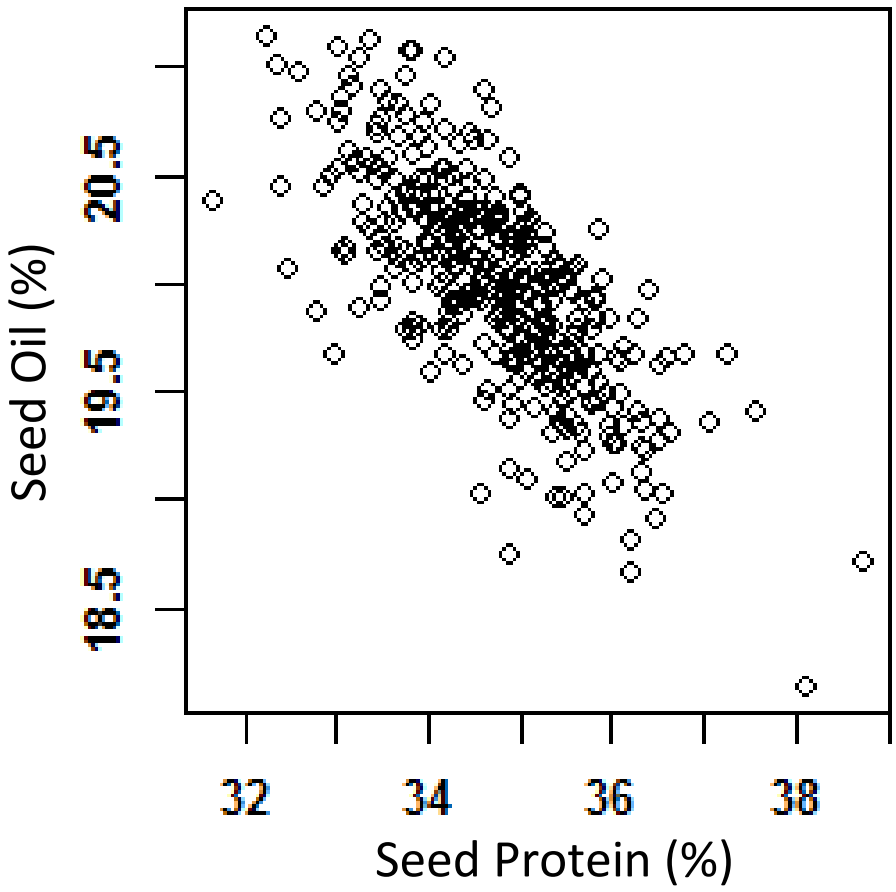
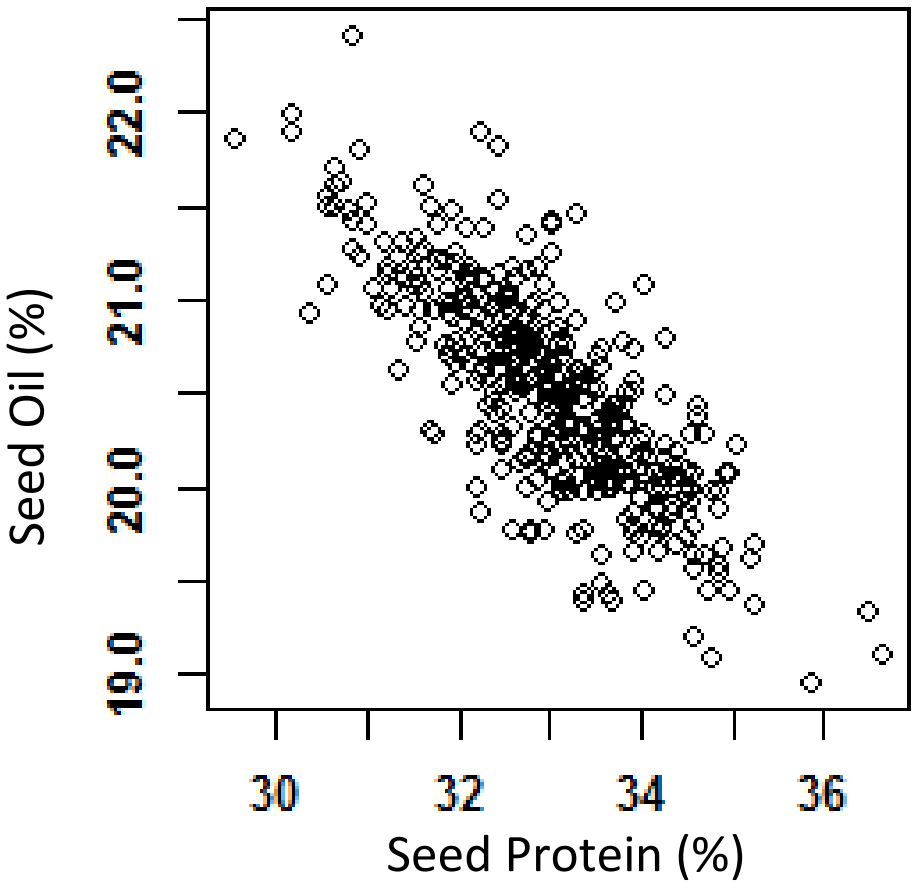


Fig. 2. The tickmarks on the vertical lines in this gragh represent the map positions of 1536 SNP markers comprising the Universal Soy Linkage Panel 1.0 (Hyten et al., 2010) within each of the 20 soybean chromosomes (top) and corresponding linkage groups (bottom). The vertical map distance is scaled in Kosambi centiMorgans.

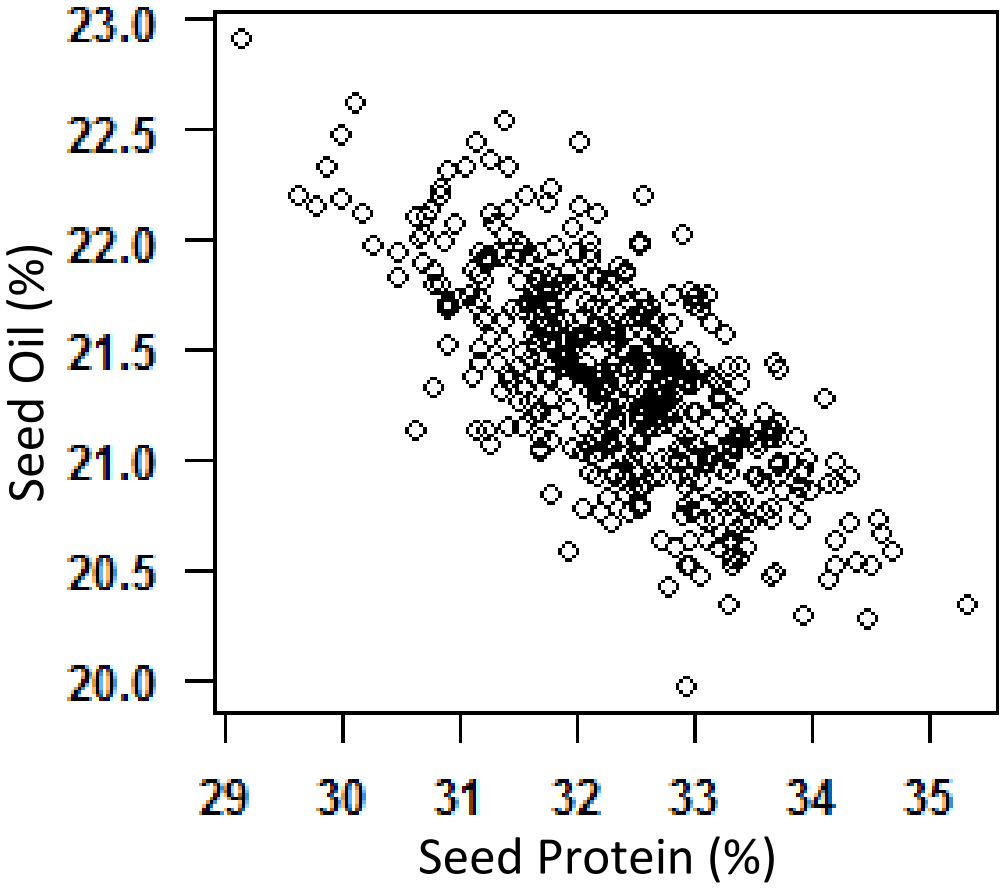
Fig. 3. A graphic illustration of seed protein and seed oil content of individual F_2 plants in (a) UX2430, (b) UX2428, and (c) UX2427 populations.



(a)

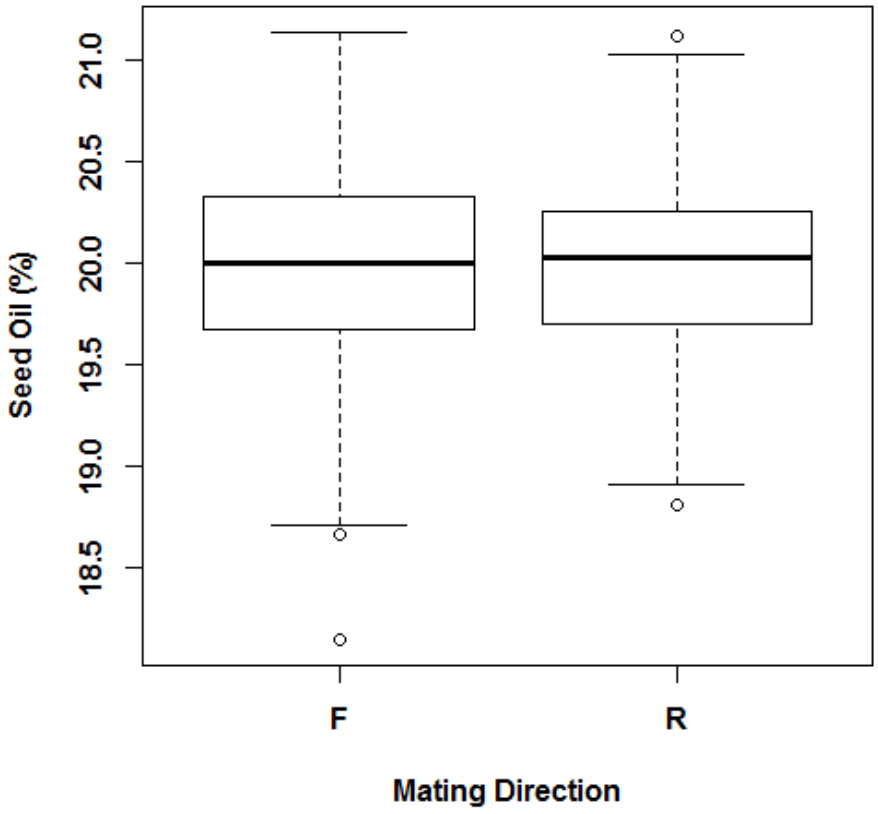


(b)

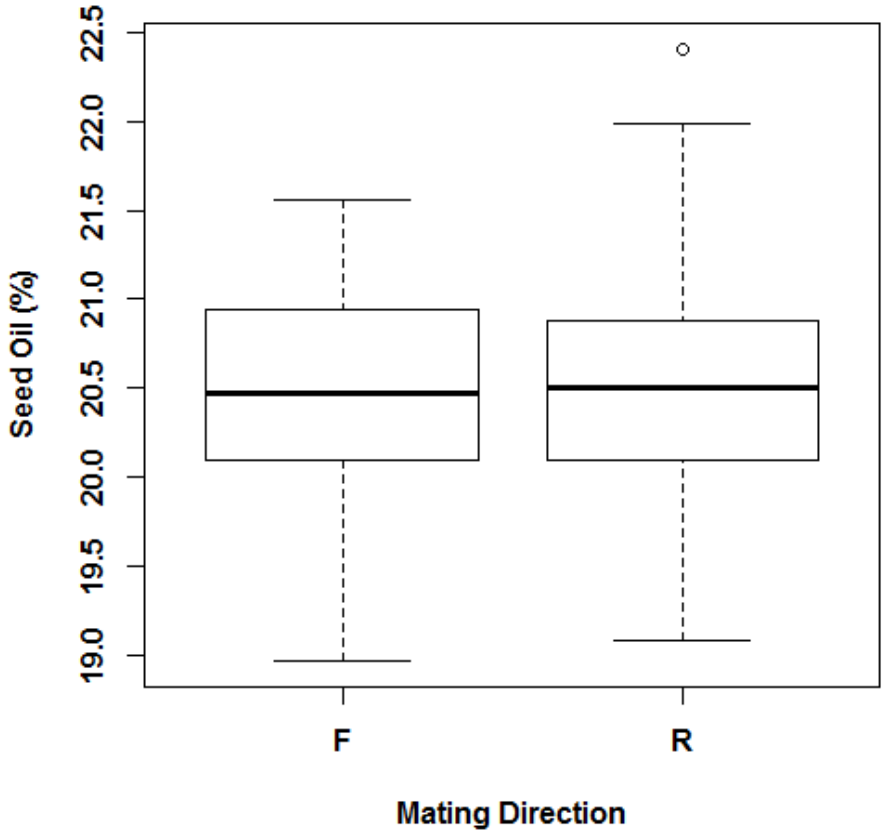

23.0
22.5
22.0
21.5
21.0
20.5
20.0
Seed Oil (%)
29
30
31
32
33
34
35
Seed Protein (%)

(c)

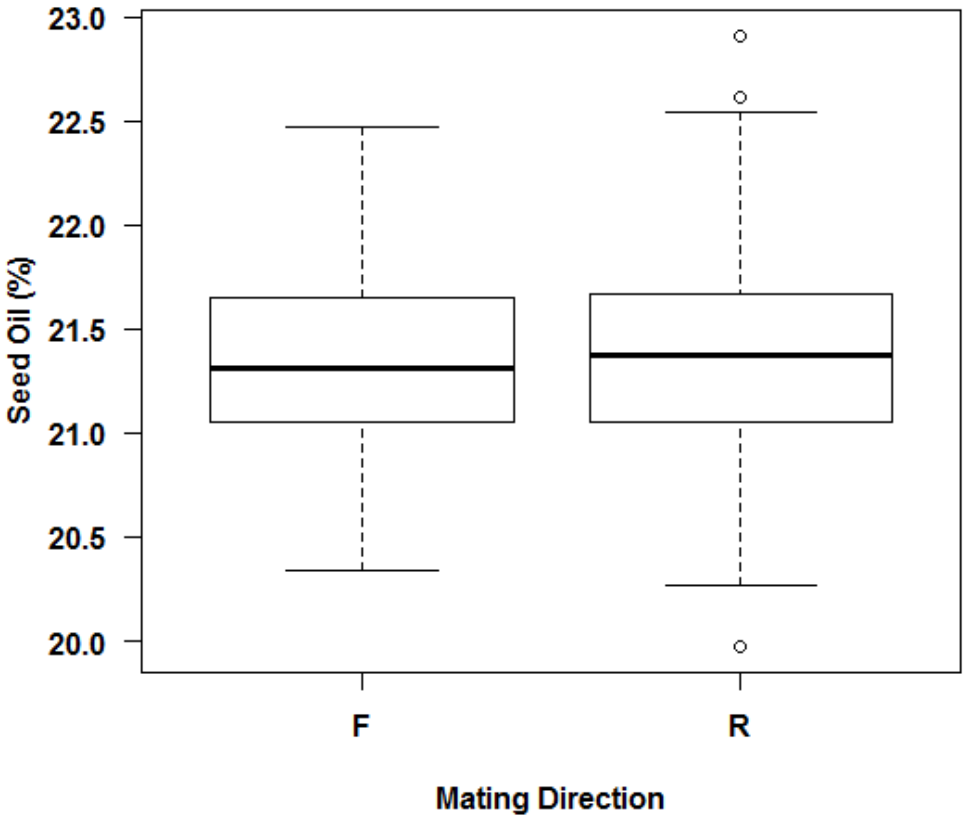
Fig. 4. Boxplots for seed oil content of two mating directions in (a) UX2430, (b) UX2428, and (c) UX2427 $F_{2.4}$ populations.



(a)

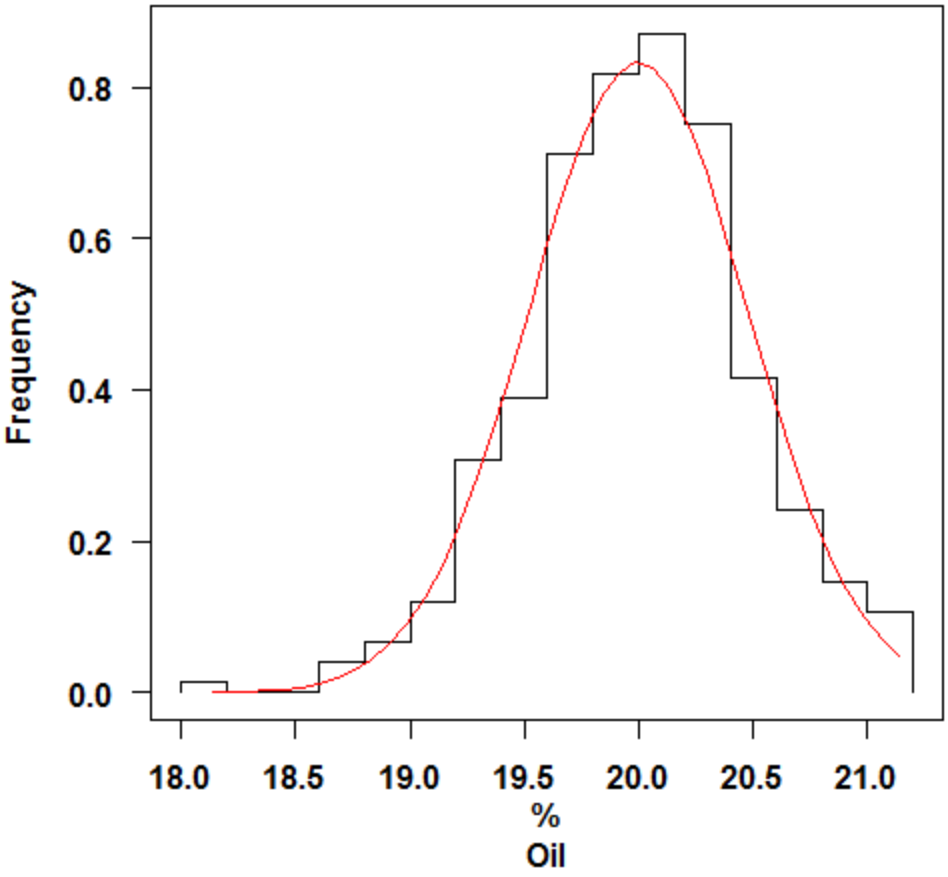


(b)

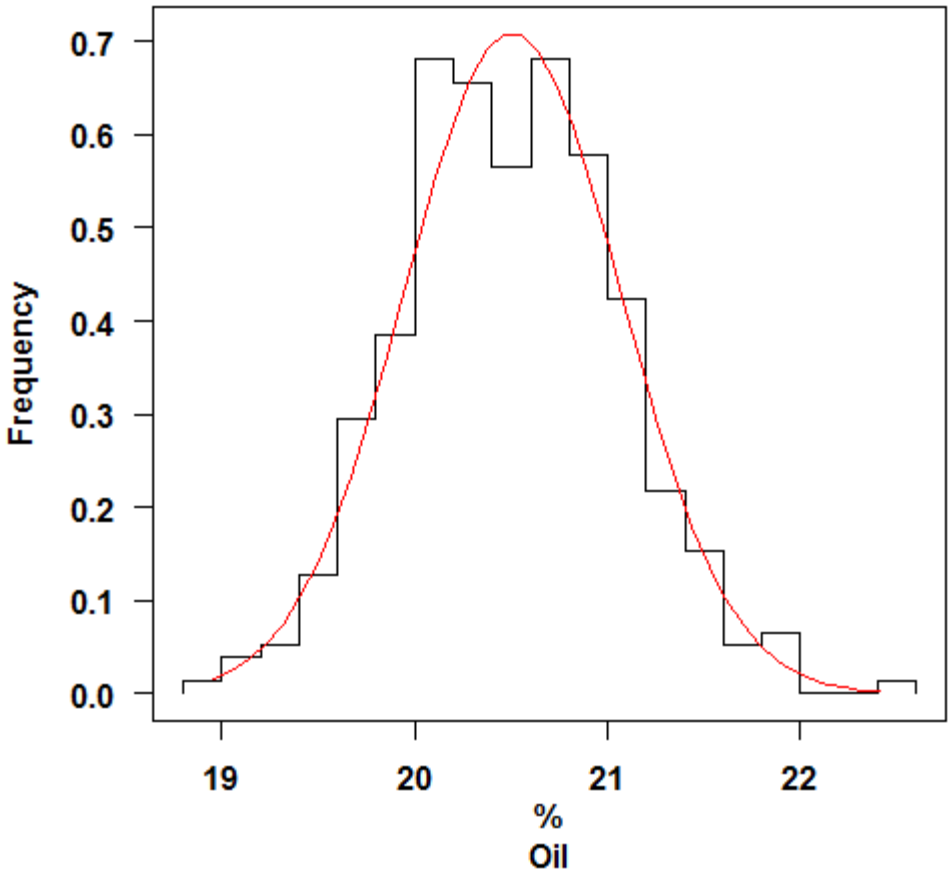


(c)

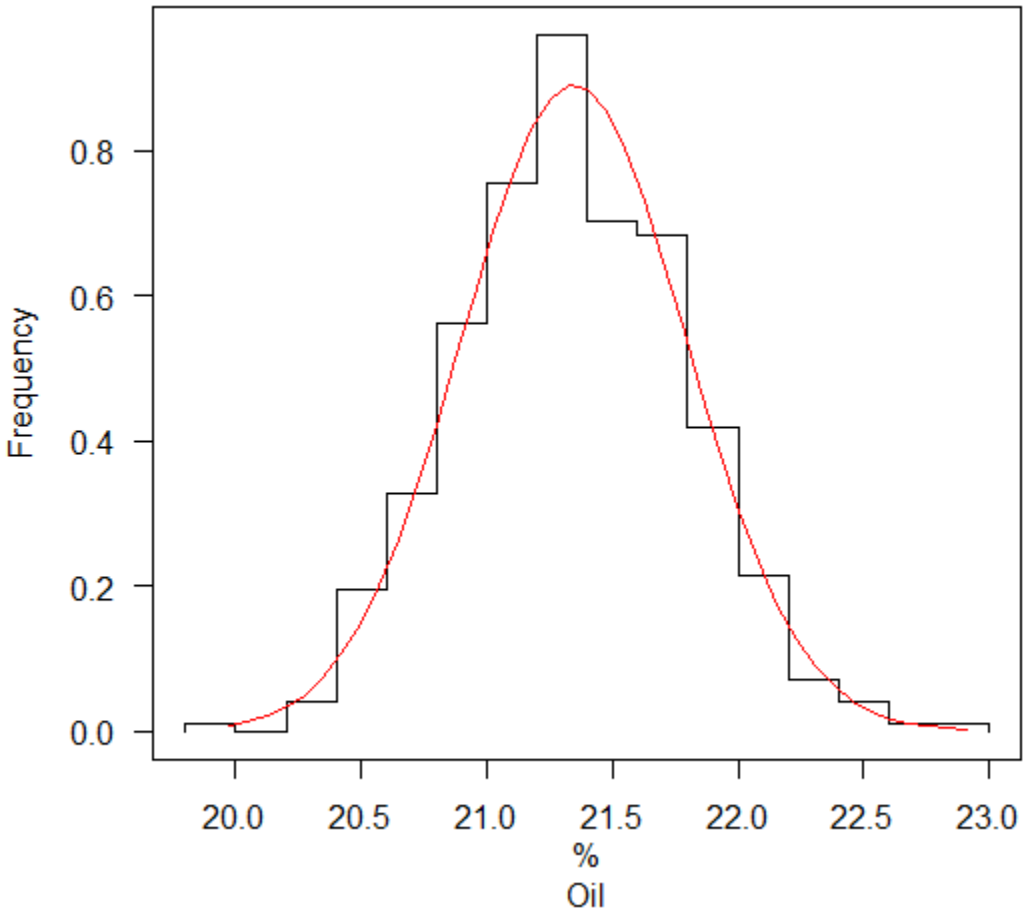
Fig. 5. Histogram distributions for seed oil phenotype in (a) UX2430, (b) UX2428, and (c) UX2427 $F_{2.4}$ populations. The solid line is showed normal distribution curve.



(a)

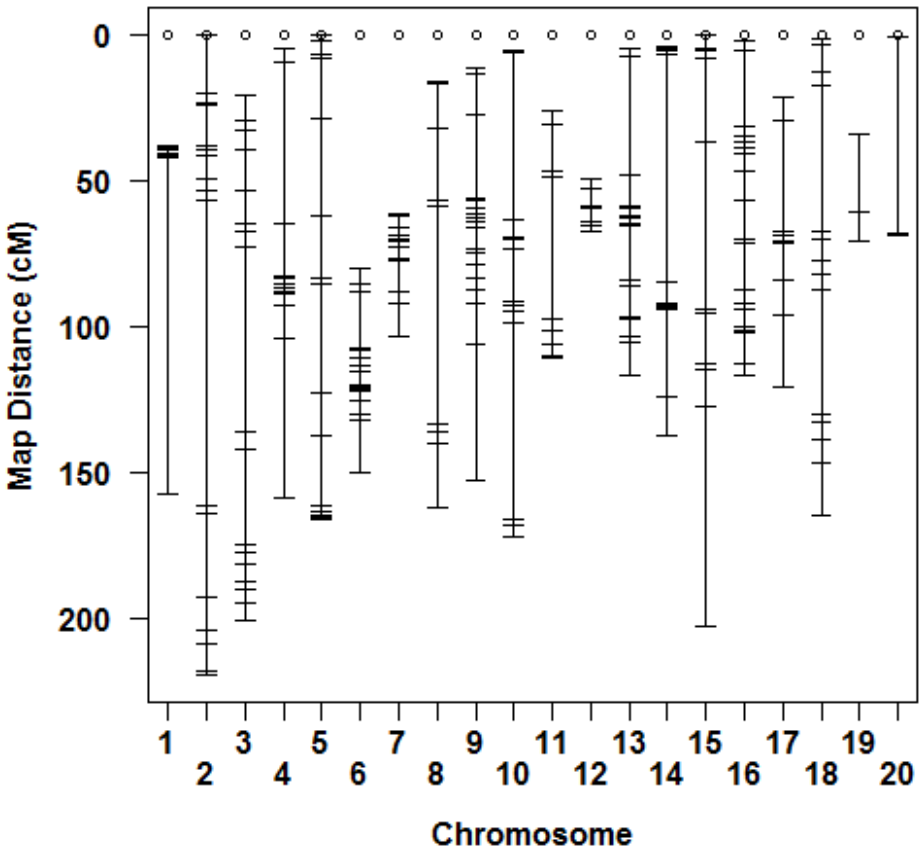


(b)

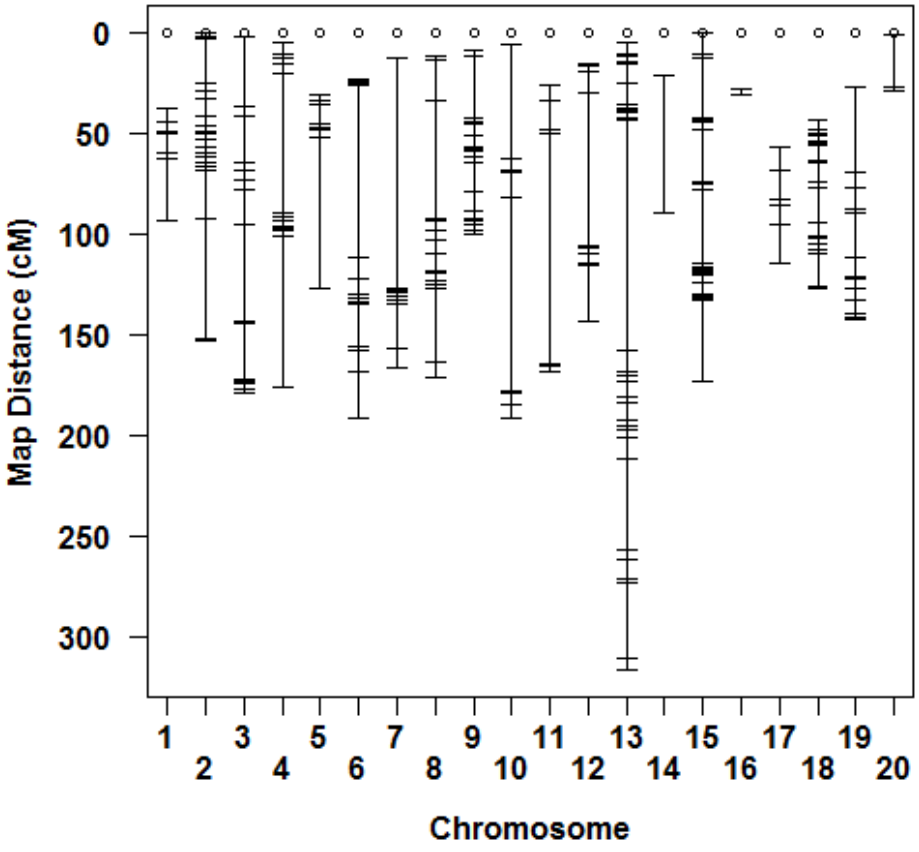


(c)

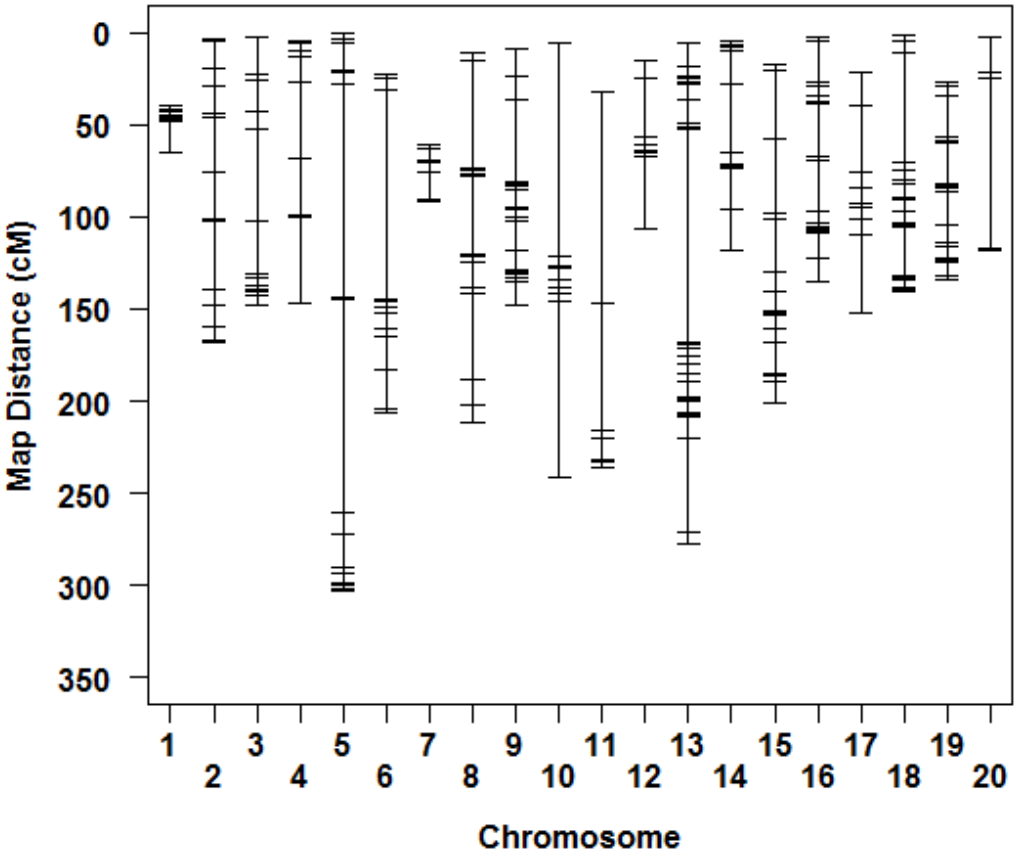
Fig. 6. The SNP marker genetic maps constructed for (a) UX2430, (b) UX2428, and (c) UX2427 F_2 populations. About 320-370 SNP markers remained in each population for final linkage map construction.


Map Distance (cM)
0
50
100
150
200
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(a)

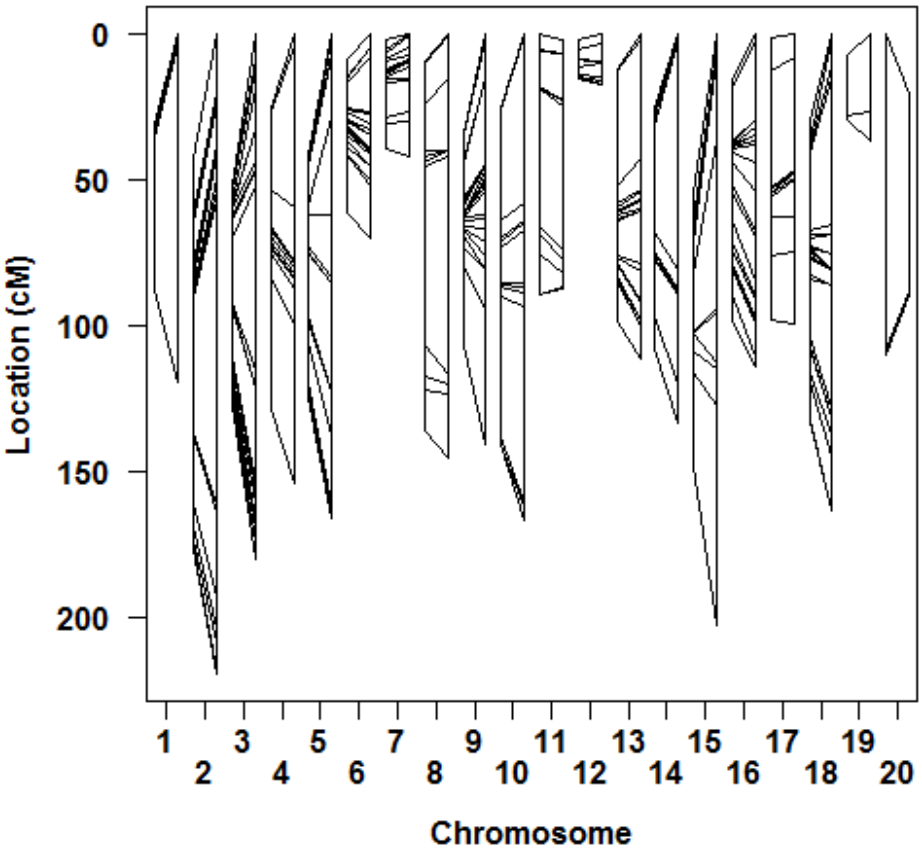

Map Distance (cM)
0
50
100
150
200
250
300
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(b)

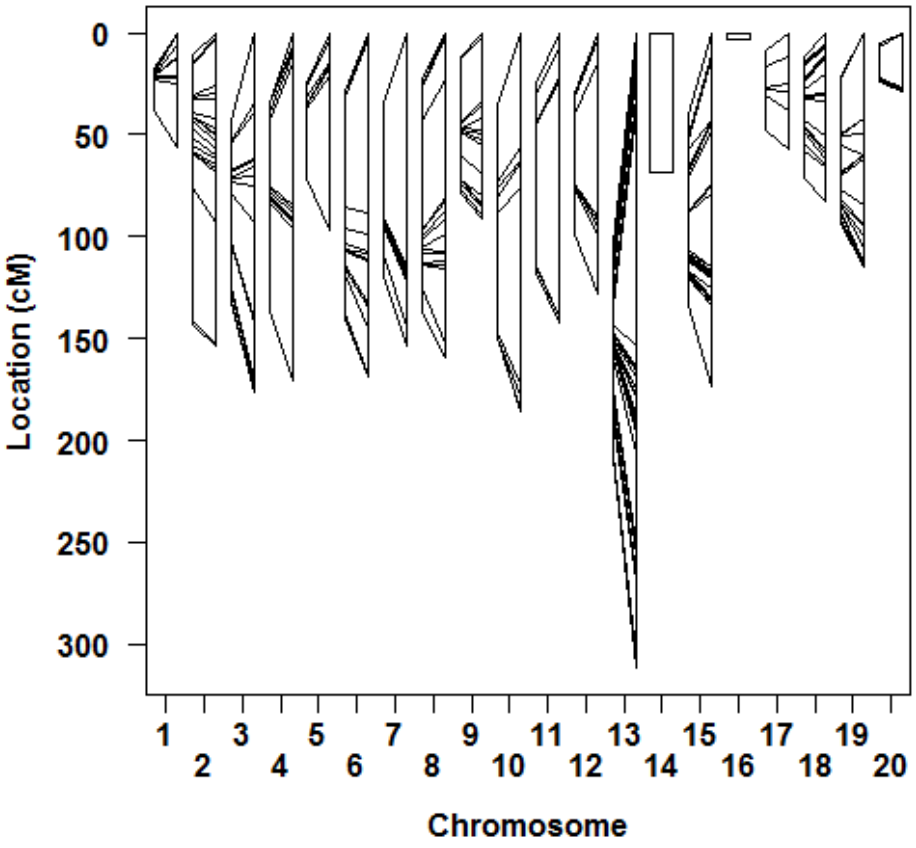

Map Distance (cM)
0
50
100
150
200
250
300
350
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(c)

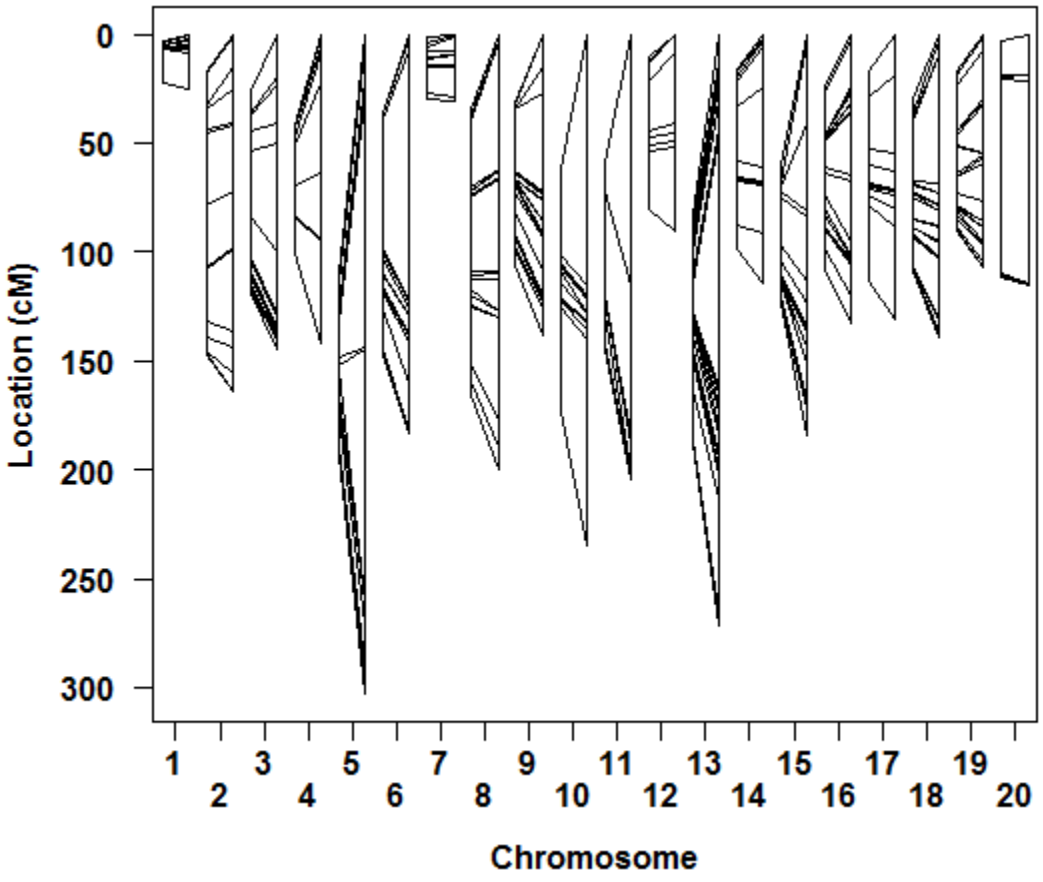
Fig. 7. Comparison of chromosomal map lengths and markers position of Hyten linkage map (left side) and (a) UX2430, (b) UX2428, and (c) UX2427 F_2 linkage map (right side).


Location (cM)
0
50
100
150
200
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(a)

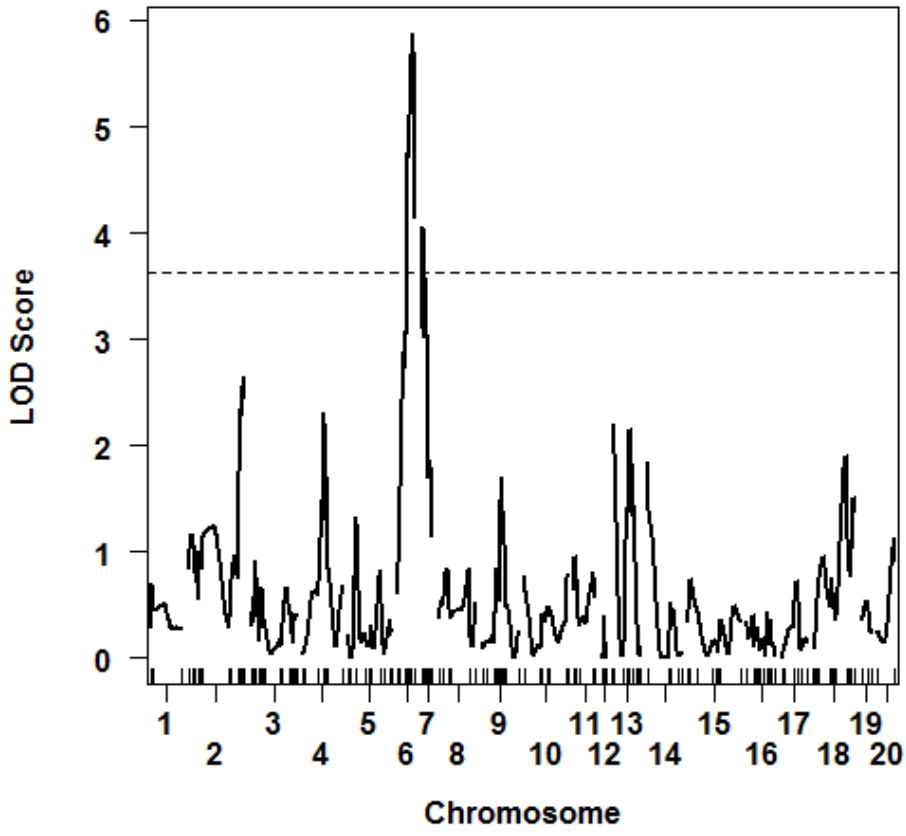

Location (cM)
0
50
100
150
200
250
300
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(b)

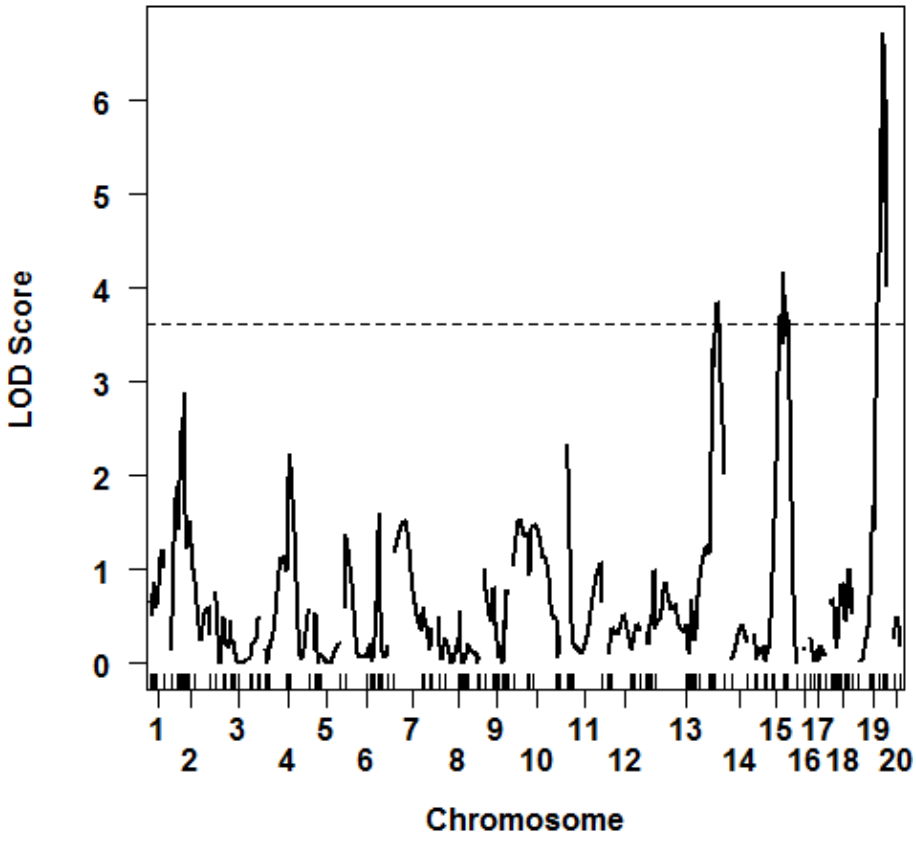

Location (cM)
0
50
100
150
200
250
300
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(c)

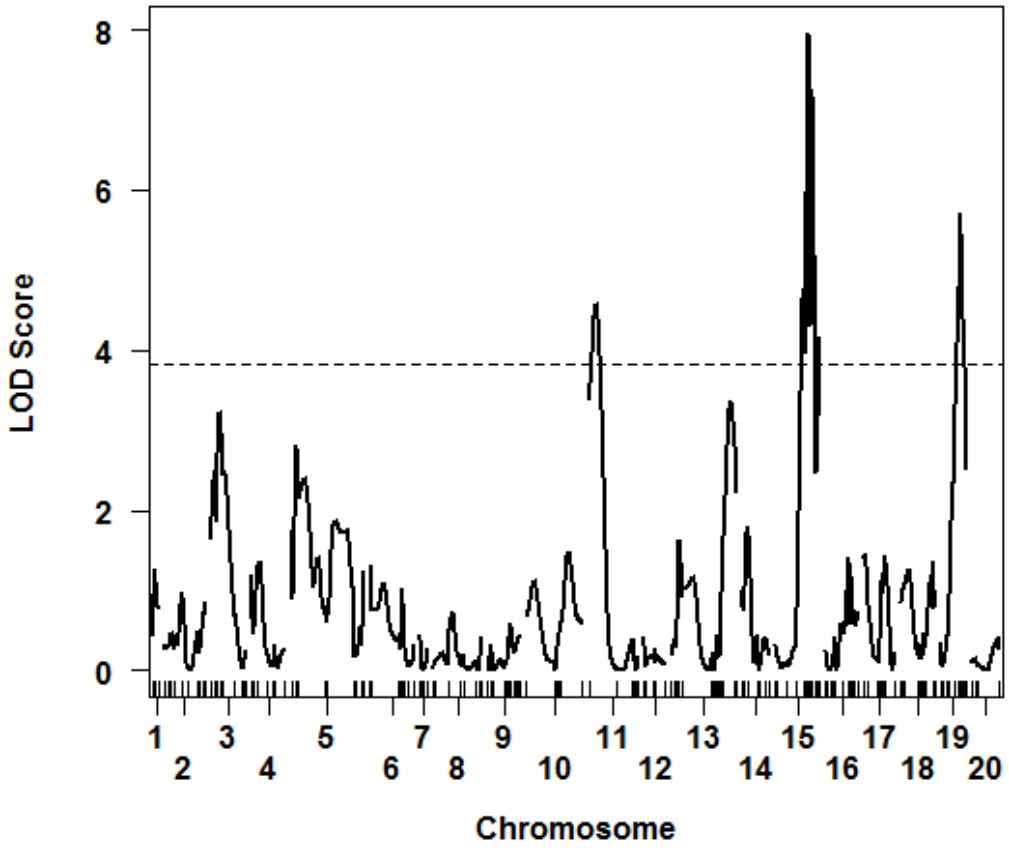
Fig. 8. Shown here are the genome-wide seed oil LOD score scans generated using the interval analysis method (i.e., maximum likelihood approach using the EM algorithm) with respect to the selectively genotyped $F_{2.4}$ progeny seed oil values in (a) UX2430, (b) UX2428, and (c) UX2427 $F_{2.4}$ populations.The LOD score for significance (dashed line) in each population was determined by using the 95_{th} percentile of genome-wide maximum LOD scores obtained from 1900 replicates of stratified permutation.


LOD Score
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(a)

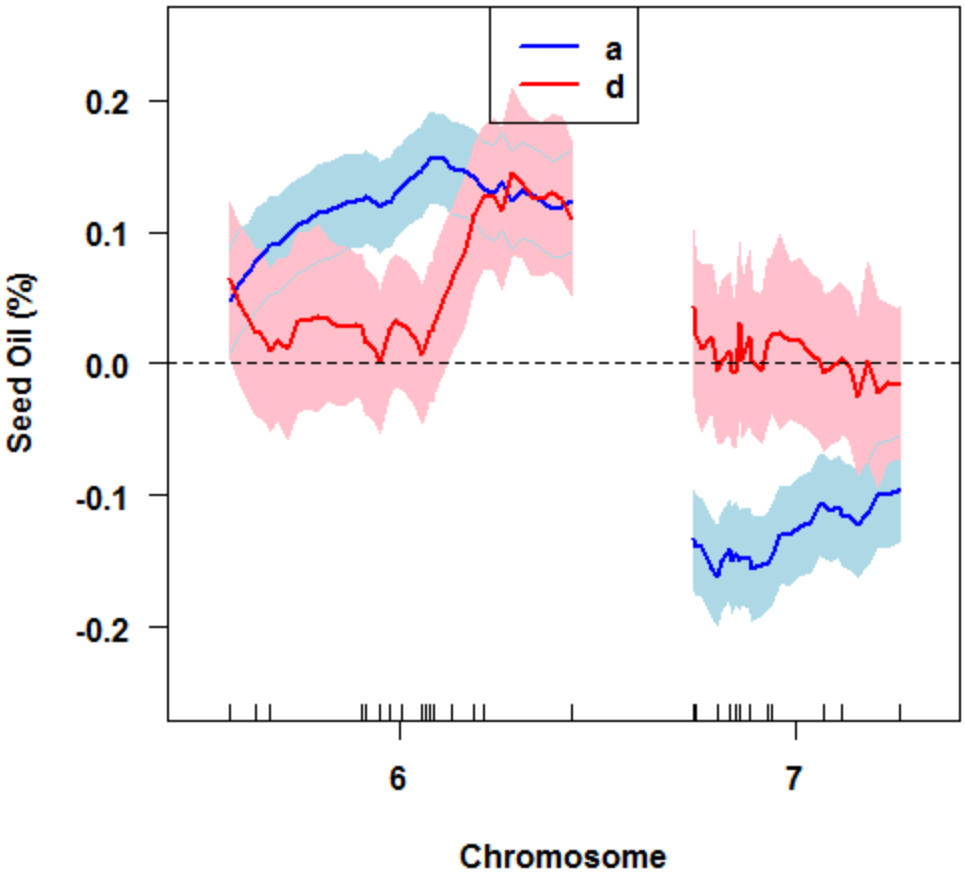

LOD Score
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(b)

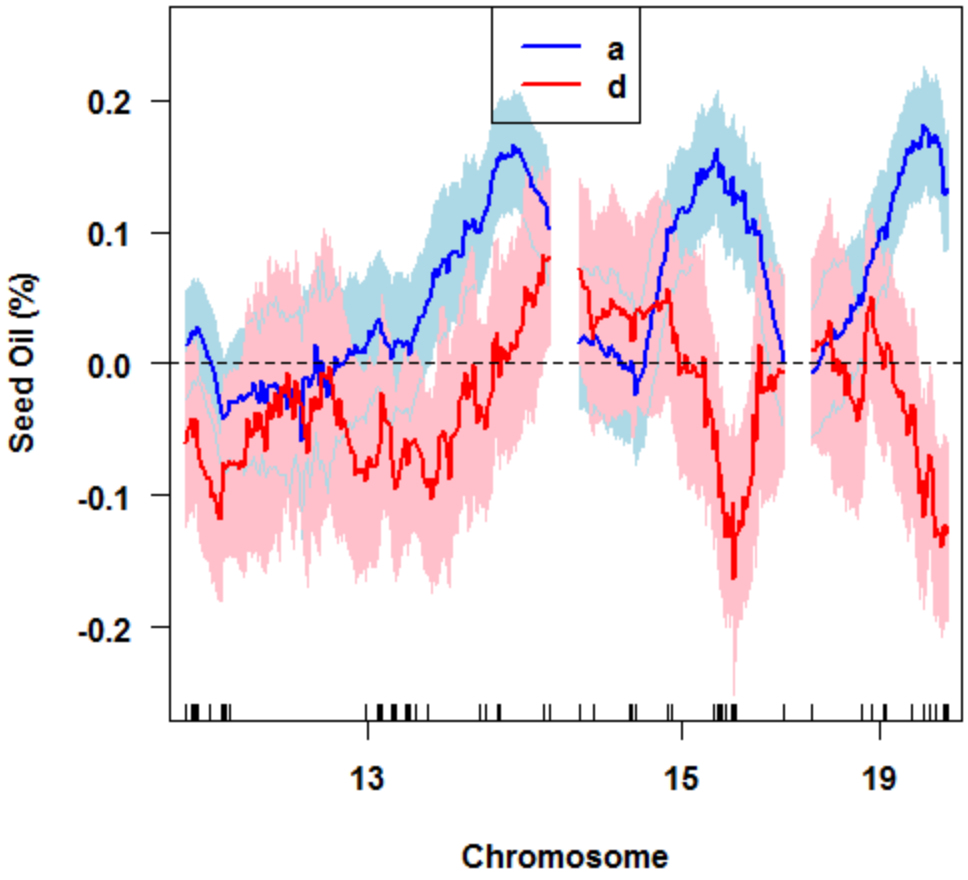

LOD Score
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(c)

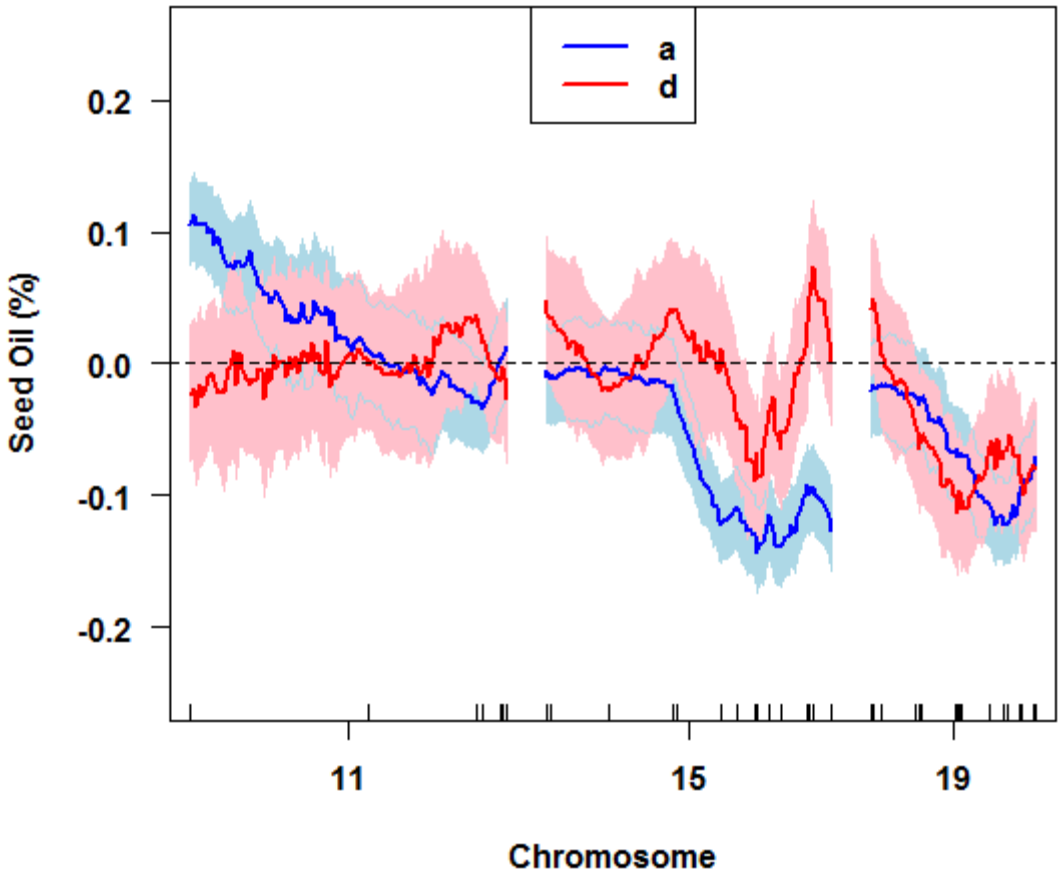
Fig. 9. Here are shown the additive (a) and dominant (d) effects on seed oil content of statistically significant alleles (only the relevant chromosomes displayed here) in (a) UX2430, (b) UX2428, and (c) UX2427 $F_{2.4}$ populations. The additive and dominant effects were estimated by linear regression of oil content phenotypes onto A/H/B genotypes.



(a)



(b)



(c)

REFERENCES

Arends, D., P. Prins, R.C. Jansen, and K.W. Broman. 2010. R/qtl: High-throughput multiple QTL mapping. Bioinformatics. 26: 2990-2992.

Ayoub, M., and D.E. Mather. 2002. Effectiveness of selective genotyping for detection of quantitative trait loci: an analysis of grain and malt quality traits in three barley populations. Genome 45:1116-1124.

Bernard, R.L. and C.R. Cremeens. 1988. Registration of 'Williams 82' soybean. Crop Sci. 28:1027-1028.

Botstein, D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using Restriction Fragment Length Polymorphisms. American Society of Human Genetics. 32:314-331.

Brim, C.A. 1973. Quantitative genetics and breeding, p. 155–186, *In* B. E. Caldwell, ed. Soybeans: Improvement, production, and uses. ASA, Madison, WI.

Brim, C.A., and J.W. Burton. 1979. Recurrent selection in soybeans. II. Selection for increased percent protein in seeds. Crop Sci. 19:494-498.

Broman, K.W., H. Wu, Ś. Sen, and G.A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890.

Brown, J. and P. Caligari. 2008. An introduction to plant breeding. Wiley-Blackwell, Iowa.

Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, and R.C. Shoemaker. 1997. Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci. 37:370-378.

Burton, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding, p. 211-

242, *In* J. R. Wilcox, ed. Soybeans: Improvement, production, and uses, 2nd ed. ASA, CSSA, and SSSA, Madison, WI.

Chapman, A., V.R. Pantalone, A. Ustun, F.L. Allen, D. Landau-Ellis, R.N. Trigiano, and P.M. Gresshoff. 2003. Quantitative trait loci for agronomic and seed quality traits in an F2 and F4:6 soybean population. Euphytica 129:387-393.

Choi, I.-Y., D.L. Hyten, L.K. Matukumalli, Q. Song, J.M. Chaky, C.V. Quigley, K. Chase, K.G. Lark, R.S. Reiter, M.-S. Yoon, E.-Y. Hwang, S.-I. Yi, N.D. Young, R.C. Shoemaker, C.P. van Tassell, J.E. Specht, and P.B. Cregan. 2007. A soybean transcript map: Gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685-696.

Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, P.B. Cregan, R.C. Shoemaker, and J.E. Specht. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci. 43:1053-1067.

Cober, E.R., and H.D. Voldeng. 2000. Developing high-protein, high-yield soybean populations and lines. Crop Sci. 40:39-42.

Collard, B., M. Jahufer, J. Brower, and E. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196.

Coryell, V.H., H. Jessen, J.M. Schupp, D. Webb, and P. Keim. 1999. Allele-specific hybridization markers for soybean. Theor. Appl. Genet. 98:690-696.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, and J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464-1490.

Csanádi, G., J. Vollmann, G. Stift, and T. Lelley. 2001. Seed quality QTLs identified in a molecular map of early maturing soybean. Theor. Appl. Genet. 103:912-919.

Darvasi, A. 1997. The effect of selective genotyping on QTL mapping accuracy. Mamm. Genome 8:67-68.

Darvasi, A., and M. Soller. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. 85:353-359.

Darvasi, A., and M. Soller. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics 138:1365-1373.

Diers, B.W., P. Keim, W.R. Fehr, and R.C. Shoemaker. 1992. RFLP analysis of soybean seed protein and oil content. Theor. Appl. Genet. 83:608-612.

Fasoula, V.A., D.K. Harris, and H.R. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. Crop Sci. 44:1218-1225.

GRIN. 2011. Germplasm resources information network - National Genetic Resources Program (NGRP). Administered online by USDA-ARS. http://www.ars-grin.gov/ (Verified 25 June 2011).

Hanson, W.D., R.C. Leffel, and R.W. Howell. 1961. Genetic analysis of energy production in the soybean. Crop Sci. 1:121–126.

Hartwig, E.E., and T.C. Kilen. 1991. Yield and composition of soybean seed from parents with different protein, similar yield. Crop Sci. 31:290-292.

Hearne, C.M., S. Ghosh, and J.A. Todd. 1992. Microsatellites for linkage analysis of genetic traits. Trends in Genetics. 8:288-294.

Helms, T.C., and J.H. Orf. 1998. Protein, oil, and yield of soybean lines selected for increased protein. Crop Sci. 38:707-711.

Hwang, T., M. Sayama, Y. Takada, Y. Nakamoto, H. Funatsuki, H. Hisano, S. Sasamoto, S. Sato, S. Tabata, I. Kono, M. Hoshi, M. Hanawa, C. Yano, Z. Xia, K. Harada, K. Kitamura, and M. Ishimoto. 2009. High-density integrated linkage map based on SSR markers in Soybean. DNA Research. 16:213-225.

Hyten, D., Q. Song, I.-Y. Choi, M.-S. Yoon, J. Specht, L. Matukumalli, R. Nelson, R. Shoemaker, N. Young, and P. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116:945-952.

Hyten, D.L., V.R. Pantalone, C.E. Sams, A.M. Saxton, D. Landau-Ellis, T.R. Stefaniak, and M.E. Schmidt. 2004. Seed quality QTL in a prominent soybean population. Theor. Appl. Genet. 109:552-561.

Hyten, D.L., Q. Song, Y. Zhu, I.-Y. Choi, R.L. Nelson, J.M. Costa, J.E. Specht, R.C. Shoemaker, and P.B. Cregan. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. USA 103:16666-16671.

Hyten, D.L., I.-Y. Choi, Q. Song, J.E. Specht, T.E. Carter, R.C. Shoemaker, E.-Y. Hwang, L.K. Matukumalli, and P.B. Cregan. 2010. A high density integrated genetic linkage map of soybean and the development of a 1536 Universal Soy Linkage Panel for quantitative trait locus mapping. Crop Sci. 50:960-968.

Kabelka, E.A., B.W. Diers, W.R. Fehr, A.R. LeRoy, I.C. Baianu, T. You, D.J. Neece, and

R.L. Nelson. 2004. Putative alleles for increased yield from soybean plant introductions. Crop Sci. 44:784-791.

Keim, P., B.W. Diers, T.C. Olson, and R.C. Shoemaker. 1990. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735-742.

Kumar, P., V.K. Gupta,A.K. Misra, D.R. Modi, and B.K. Pandey. Potential of molecular markers in plant biotechnology. Plant Omics Journal. 2009. l2(4):141-162.

Lander, E.S., and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.

Lark, K.G., J.M. Weisemann, B.F. Matthews, R. Palmer, K. Chase, and T. Macalma. 1993. A genetic map of soybean (*Glycine max* L.) using an intraspecific cross of two cultivars: 'Minsoy' and 'Noir 1'. Theor. Appl. Genet. 86:901-906.

Lark, K.G., J. Orf, and L.M. Mansur. 1994. Epistatic expression of quantitative trait loci (QTL) in soybean [*Glycine max* (L.) Merr.] determined by QTL association with RFLP alleles. Theor. Appl. Genet. 88:486-489.

Lebowitz, R.J., M. Soller, and J.S. Beckmann. 1987. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet. 73:556-562.

Lee, S.H., M.A. Bailey, M.A.R. Mian, T.E. Carter, E.R. Shipe, D.A. Ashley, W.A. Parrott, R.S. Hussey, and H.R. Boerma. 1996. RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor. Appl. Genet. 93:649-657.

Mansur, L.M., K.G. Lark, H. Kross, and A. Oliveira. 1993a. Interval mapping of

quantitative trait loci for reproductive, morphological, and seed traits of soybean (Glycine max L.). Theor. Appl. Genet. 86:907-913.

Mansur, L.M., J. Orf, and K.G. Lark. 1993b. Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (*Glycine max* L. Merr.). Theor. Appl. Genet. 86:914-918.

Mansur, L.M., J.H. Orf, K. Chase, T. Jarvik, P.B. Cregan, and K.G. Lark. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci. 36:1327-1336.

Monteros, M.J., J.W. Burton, and H.R. Boerma. 2008. Molecular mapping and confirmation of QTLs associated with oleic acid content in N00-3350 soybean. Crop Sci. 48:2223-2234.

Neale, D.B. and C.G. Williams. 1991. Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. Canadian Journal of Forest Research. 21:545-554.

Orf, J.H., K. Chase, F.R. Adler, L.M. Mansur, and K.G. Lark. 1999. Genetics of soybean agronomic traits: II. Interactions between yield quantitative trait loci in soybean. Crop Sci. 39:1652-1657.

Palmer, R.G., T.W. Pfeiffer, G.R. Buss and T.C. Kilen. 2004. Qualitative genetics. p. 137–233. In J.R. Wilcox, H.R. Boerma, and J.E. Specht (ed.) Soybeans: Improvement, production, and uses. Agronomy Monogr. 16. 3rd ed. ASA, CSSA, and SSSA, Madison, WI.

Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, and C.E. Sams. 2005. Quantitative trait loci for seed protein and oil concentration, and seed size in

soybean. Crop Sci. 45:2015-2022.

Panthee, D.R., V.R. Pantalone, and A. M. Saxton. 2006. Modifier QTL for fatty acid composition in soybean oil. Euphytica. 152:67-73.

Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding. 2:225-238.

Qiu, B.X., P.R. Arelli, and D.A. Sleper. 1999. RFLP markers associated with soybean cyst nematode resistance and seed composition in a 'Peking'×'Essex' population. Theor. Appl. Genet. 98:356-364.

Qi, Z., Q. Wu, X. Han, Y. Sun, X. Du, C. Liu, H. Jiang, G. Hu, and Q. Chen. 2011. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica. 179:499-514.

Ramteke R., V. Kumar, P. Murlidharan, and D.K. Agarwal. 2010. Study on genetic variability and traits interrelationship among released soybean varieties of India [Glycine max (L.) Merrill]. Electronic Journal of Plant Breeding. 1(6):1483-1487.

Ritchie, R.A. 2003. High-protein plant introductions: Selective genotyping to detect soybean protein QTLs. M.S. thesis, Univ. of Nebraska, Lincoln.

Sebern, N.A., and J.W. Lambert. 1984. Effect of stratification for percent protein in two soybean populations. Crop Sci. 24:225–228.

Sebolt, A.M., R.C. Shoemaker, and B.W. Diers. 2000. Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci. 40:1438-1444.

Schwender, J., J.B. Ohlrogge, and Y. Shachar Hill. 2003. A flux model of glycolysis and the oxidative pentosephosphate pathway in developing Brassica napus embryos. J. Biol. Chem. 278:29442–29453.

Shannon, J.G., J.R. Wilcox, and A.H. Probst. 1972. Estimated gains from selection for protein and yield in the F4 generation of six soybean populations. Crop Sci. 12:824–826.

Shoemaker, R.C., and T.C. Olson. 1993. Molecular linkage map of soybean (*Glycine max* L. Merr.), p. 6131-6138, *In* S. J. O'Brien, ed. Genetic maps: locus maps of complex genomes. Cold Spring Harbor Laboratory Press, NY.

Shoemaker, R.C., and J.E. Specht. 1995. Integration of the soybean molecular and classical genetic linkage groups. Crop Sci. 35:436-446.

Smith D.N. and M.E. Devey. 1994. Occurrence and inheritance of microsatellite loci in Pinus radiata. Genome. 37:977–983.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, J.E. Specht, and P.B. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109:122-128.

Soybase. 2011. SoyBase– a genome database for Glycine. Administered online by USDA-ARS and Iowa State University. http://soybeanbreederstoolbox.org/ (Verified 25 June 2011).

Soystats. 2011. Soy Stats: A reference guide to important soybean fact & figures. http://www.soystats.com/2011/Default-frames.htm (Verified 25 June 2011).

Specht, J.E., K. Chase, M. Macrander, G.L. Graef, J. Chung, J.P. Markwell, M. Germann, J.H. Orf, and K.G. Lark. 2001. Soybean response to water: A QTL analysis of

drought tolerance. Crop Sci. 41:493-509.

Takahashi, R., H. Matsumura, M.E. Oyoo, and N.A. Khan. 2008. Genetic and linkage analysis of purple – blue flower in soybean. Journal of Heredity. 99(6):593–597.

Thorne, J.C., and W.R. Fehr. 1970. Incorporation of high-protein, exotic germplasm into soybean populations by 2- and 3-way crosses. Crop Sci. 10:652-655.

Van, K., E.-Y. Hwang, M.Y. Kim, Y.-H. Kim, Y.-I. Cho, P.B. Cregan, and S.-H. Lee. 2004. Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. Euphytica 139:147-157.

Wehrmann, V.K., W.R. Fehr, S.R. Cianzio, and J.F. Cavins. 1987. Transfer of high seed protein to high-yielding soybean cultivars. Crop Sci. 27:927-931.

Yaklich, R.W., B. Vinyard, M. Camp, and S. Douglass. 2002. Analysis of seed protein and oil from soybean Northern and Southern Region Uniform Tests. Crop Sci. 42:1504-1515.

Zhu, T., L. Shi, J.J. Doyle, and P. Keim. 1995. A single nuclear locus phylogeny of soybean based on DNA sequence. Theor. Appl. Genet. 90:991-999.

Zhu, Y.L., Q.J. Song, D.L. Hyten, C.P. Van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young, and P.B. Cregan. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163:1123-1134.

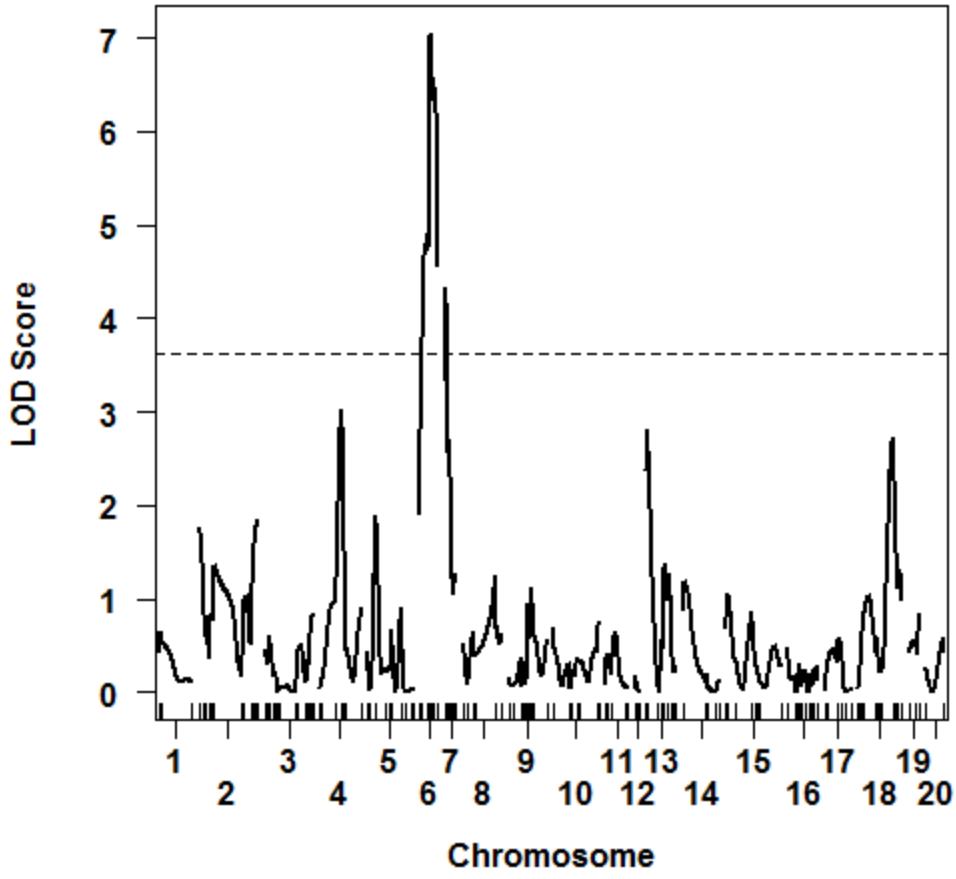
Appendix Table 1. Summary of seed protein QTL peak scores ≥ 3.0, ordered by population, then by chromosome, that were identified by interval mapping using expectation maximization (EM). A permutation test of 1900 replications was conducted in each population to provide a genome-wide 95th percentile LOD score to serve as a statistical significance criterion for evaluating a QTL LOD score peak. The additive (a) and dominant (d) effects were calculated on the basis of the substitution of a high oil low protein parent allele for a low oil high protein parent allele at the relative marker locus.

Pop.		Chr.	LG		LOD	Permutation-based		QTL Effect	
No.	SNP	No.	name	Position	(if ≥ 3.0)	LOD Score	R^2	a‡	d‡
				cM			%	---g kg-1---	
UX2427	S05979	3	N	48.7	3.16	-	-	-	-
UX2427	S06956†	11	B1	42.1	4.87	3.71	4.5	2.4	0.6
UX2427	S10061	15	E	153.3	7.31	3.71	6.6	-3.0	1.4
UX2427	S02534†	19	LG	109.4	5.02	3.71	4.6	-2.1	2.4
UX2428	S18270	2	D1b	57.2	3.25	-	-	-	-
UX2428	S17881	6	C2	158.0	4.21	3.62	4.9	-3.2	1.3
UX2428	S12241†	11	B1	30.2	3.33	-	-	-	-
UX2428	S00135	19	L	132.9	4.40	3.62	5.1	-2.9	0.5
UX2430	S10820	4	C1	88.2	3.04	-	-	-	-
UX2430	S16994†	6	C2	122.2	7.06	3.57	8.3	-4.0	0.4
UX2430	S10452†	7	M	65.3	4.32	3.57	5.2	3.2	0.6

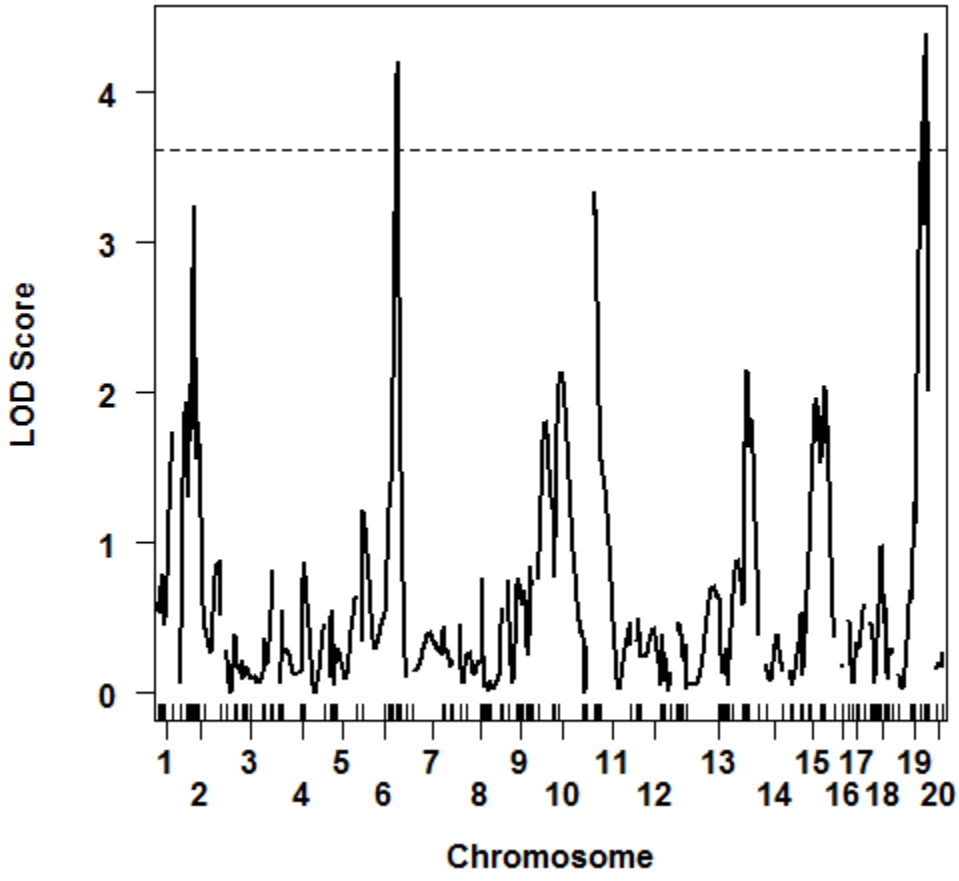
† nearest marker.

‡ If the effect is negative, the high protein parent marker allele increases seed protein content.

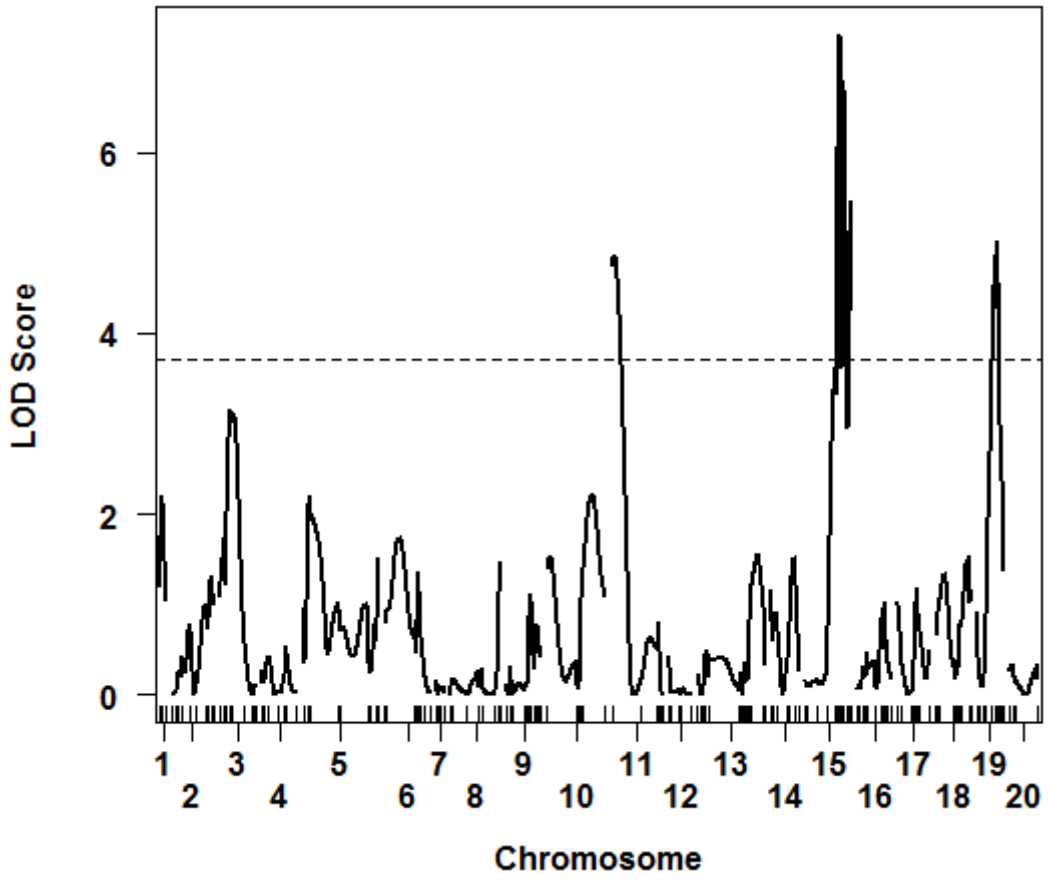
Appendix Fig. 1. Shown here are the genome-wide seed protein LOD score scans generated using the interval analysis method (i.e., maximum likelihood approach using the EM algorithm) with respect to the selectively genotyped $F_{2.4}$ progeny seed protein values in (a) UX2430, (b) UX2428, and (c) UX2427 $F_{2.4}$ populations. The LOD score for significance (dashed line) in each population was determined by using the 95^{th} percentile of genome-wide maximum LOD scores obtained from 1900 replicates of stratified permutation.


LOD Score
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(a)

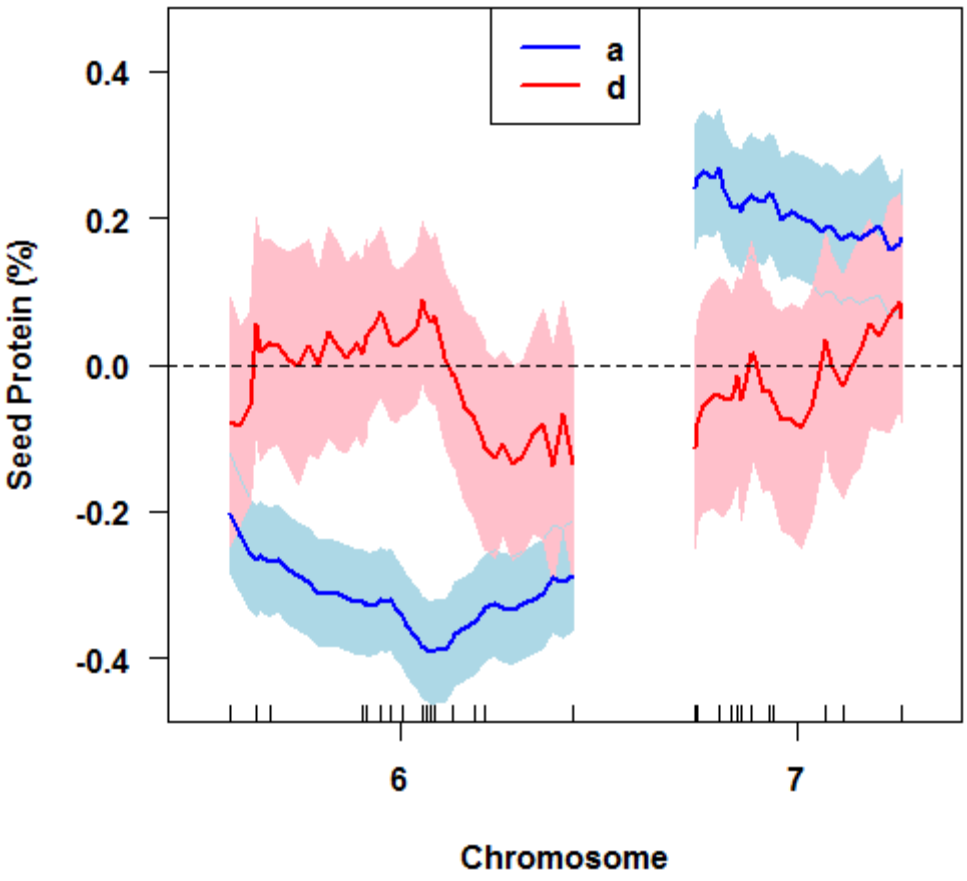

LOD Score
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(b)

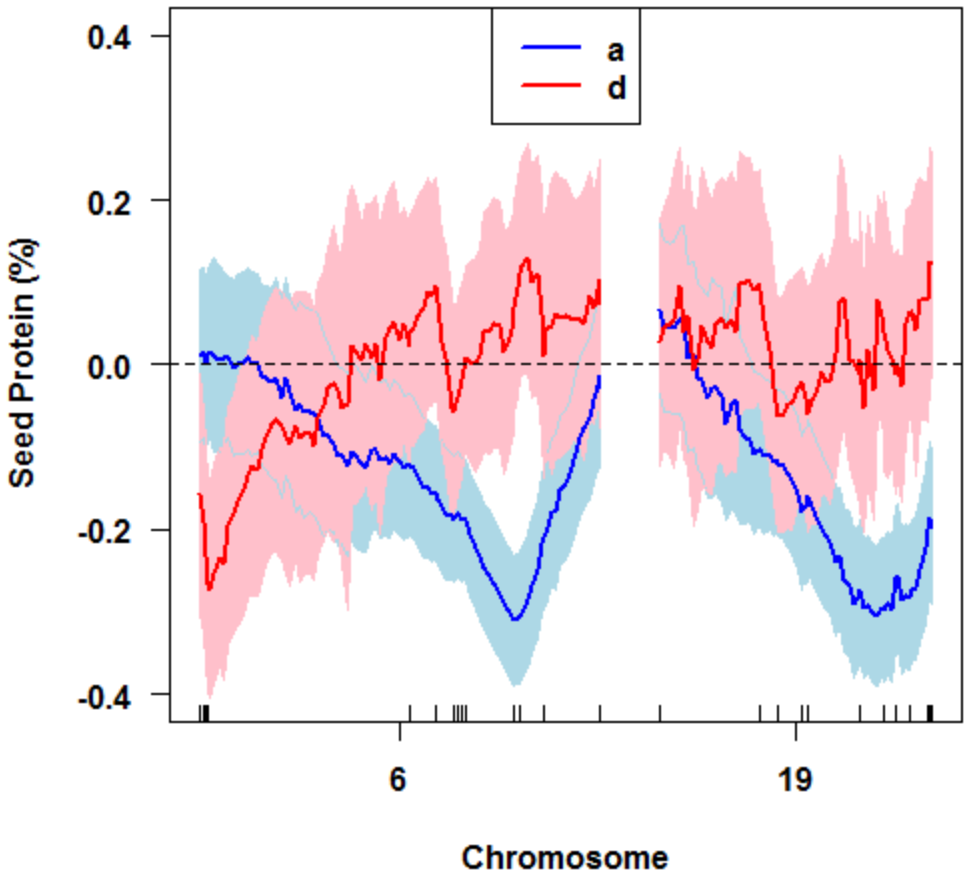

LOD Score
0
2
4
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Chromosome

(c)

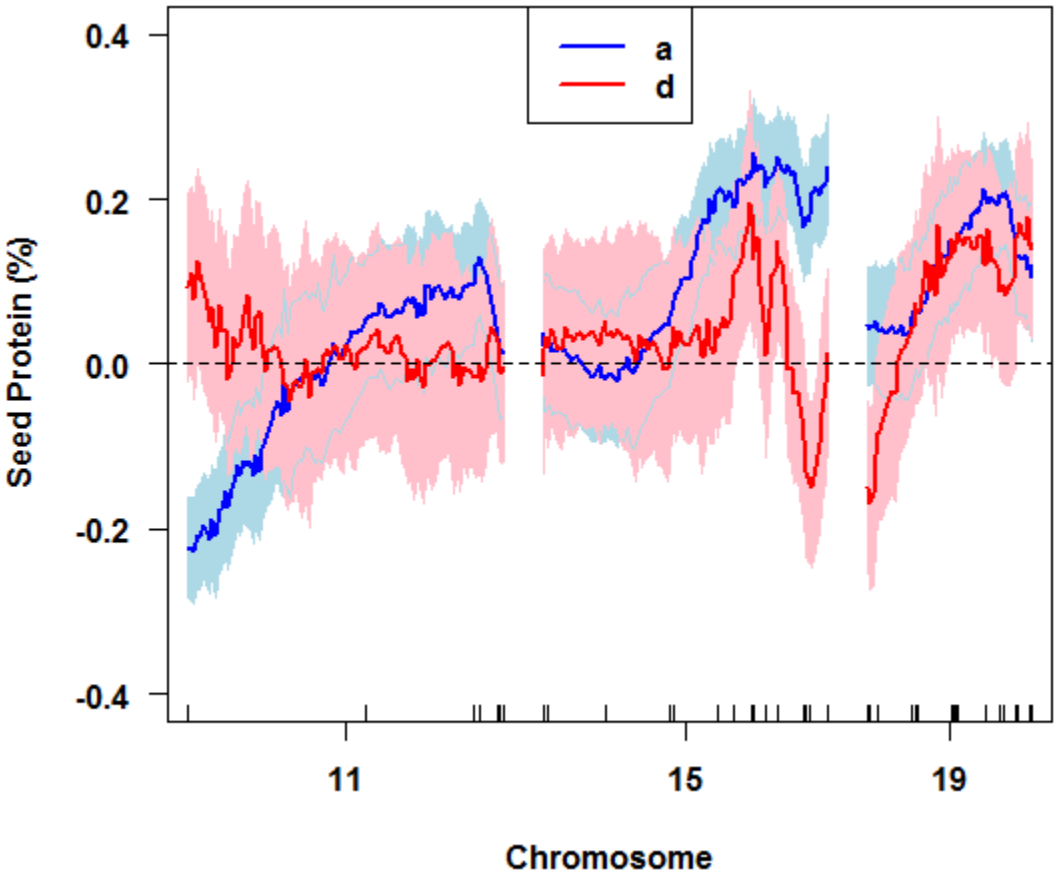
Appendix Fig. 2. Here are shown the additive (a) and dominant (d) effects on seed protein content of statistically significant alleles (only the relevant chromosomes displayed here) in (a) UX2430, (b) UX2428, and (c) UX2427 $F_{2.4}$ populations. The additive and dominant effects were estimated by linear regression of oil content phenotypes onto A/H/B genotypes.



(a)



(b)



(c)